Analiza i wizualizacja danych

Wykład 1. Wstępne informacje o analizie danych

dr Michał Sochański

- dr Michał Sochański
  - dyżury: wtorek 9.00-10.00 oraz czwartek 13.30-14.30; po zakończoneniu wszystkich siedmiu wykładów również w godzinach wykładu – wtorek 17.45-19.15
  - email: michal.sochanski@amu.edu.pl

Główne cele zajęć

- Zapoznanie się z podstawowymi zagadnieniami związanymi z analizą danych.
- Nabycie podstawowych technicznych umiejętności związanych z:
  - wczytywaniem,
  - rozpoznaniem struktury,
  - czyszczeniem,
  - przekształcaniem,
  - analizą,
  - wizualizacją,
  - raportowaniem

  danych.
- Narzędziem, z którego będziemy korzystać jest Python.

Tematyka poszczególnych wykładów

1. Wstępne informacje o analizie danych.
2. Wprowadzanie do NumPy i pandas. Wczytywanie i wstępna eksploracja danych.
3. Czyszczenie i wstępne przekształcanie danych.
4. Podsumowania i agregacje danych oraz wybrane metody statystyczne.
5. Wizualizacja danych 1.
6. Wizualizacja danych 2.
7. Raportowanie.

Struktura dzisiejszego wykładu

1. Omówienie zasad zaliczenia oraz organizacji zajęć.
2. Uwagi o używanych (i nieużywanych) narzędziach.
3. Omówienie literatury.
4. Uwagi o znaczeniu analizy danych dla współczesnej nauki, biznesu oraz życia społecznego.
5. Podstawowe pojęcia związane z analizą danych.
6. Uwagi o strukturze danych z którymi będziemy mieli do czynienia i formatach plików z danymi.

1. Organizacja zajęć i zasady zaliczenia

- Zajęcia odbywają się w b-learningu (blended-learning).
- Odbędzie się 7 wykładów. W kolejnych tygodniach w godzinach wykładu będę odbywał dyżur (w MS Teams).
- Regularny dyżur będzie się odbywał we wtorki w godz. 9.00-10.00 oraz czwartki 13.30 - 14.30.
- W związku z tym, że w ramach tego przedmiotu nie są przewidziane ćwiczenia, na wykładach będą się pojawiały elementy „praktyczne”.
- Z tego samego względu studenci są gorąco zachęcani do wykonywania ćwiczeń (załączonych na końcu slajdów z wykładów).

DataCamp

- Platforma on-line do nauki analizy danych z wykorzystaniem języków Python, R oraz innych technologii. Jest tu dostępnych ponad 300 kursów, które są bardzo popularne wśród analityków danych.

www.datacamp.com

Będziecie mieli Państwo darmowy dostęp do tego portalu przez najbliższych 6 miesięcy. Zachęcam do korzystania i poszerzania swojej wiedzy oraz zakresu umiejętności!

Zasady zaliczenia

- Zaliczenie uzyskuje się na podstawie przygotowanego projektu – raportu zawierającego analizy i wizualizacje wybranego wcześniej zbioru danych.
- Projekt będzie przygotowywany w grupach 3-osobowych.
- Terminy:
  - ustalenie składów grup – do 23 listopada
  - ustalenie zbiorów danych – do 21 grudnia
  - przesłanie raportów – do 18 stycznia
- Raporty będzie można zaliczać na dyżurach.
- Ustalanie zbiorów danych odbywać się będzie (planowo) „na żywo”, za wyjątkiem dyżuru w godzinach wykładu (17.45-19.15).
- Zaliczanie zbiorów danych odbywać się będzie „na żywo”.

Wstępne informacje o projekcie

- „Zatwierdzenie” zbiorów danych polega na potwierdzeniu przez prowadzącego, że spełniają one odpowiednie wymogi.
- Zbiory są zazwyczaj pobierane z Internetu, najczęściej z platformy www.kaggle.com
- Do tematu warunków zaliczenia wrócimy jeszcze na ostatnich zajęciach!

MS Teams

- W zakładce „Pliki” na kanale ogólnym dostępne będą następujące materiały:
  - slajdy z wykładów (wraz z ćwiczeniami) – w katalogu „Materiały z zajęć”
  - skrypty Python – w katalogu „Materiały z zajęć”
  - plik („Zapisy.xlsx”), w którym będziecie Państwo podawać składy grup oraz zapisywać się na termin zatwierdzania zbiorów i zaliczenia projektów
- Dla każdego zespołu stworzony zostanie osobny kanał, w którym będzie można zostawić raport, zostawiać wiadomości na czacie albo odbyć rozmowę (np. w celu ustalenia zbioru danych).
- Wiadomości do całego roku będę wysyłał mailem z USOSa.

2. Uwagi o używanych (i nieużywanych) narzędziach

Mam zbiór danych – jakiego narzędzia użyć do jego analizy?

Dlaczego Python?

- Jako oprogramowania typu *open-source,* Python jest darmowy.
- Bardzo duża społeczność użytkowników wciąż tworzy nowe pakiety umożliwiające lub ułatwiające różne typy analiz.
- Powtarzalność wykonywanych analiz i przekształceń - skrypty, z których skorzystaliśmy mogą być użyte wielokrotnie.
- Python jest szybki (szybszy niż np. Excel), jeśli chodzi o ładowanie i analizę danych, co jest kluczowe w przypadku dużych zbiorów danych.
- Python jest też o wiele bardziej praktyczny przy dużych projektach, wymagających integracji dużych ilości danych z różnych źródeł i automatyzacji ich przetwarzania, w celu np. utrzymywania jakiejś aplikacji albo automatyzacji raportowania.
- Możliwość implementowania algorytmów/ metod uczenia maszynowego (*machine learning*).

Kilka uwag o Excelu

- W praktyce, Excel jest w jakimś stopniu używany przez wszystkich analityków danych, jest więc ważnym narzędziem dla każdego analityka danych.
- Excel ma wiele zalet, w szczególności:
  - Łatwy w obsłudze graficzny *interface*.
  - Wygodny, bezpośredni dostęp do tabel, który daje wrażenie bardziej „bezpośredniego kontaktu z danymi”. Dane można tu oglądać za pomocą myszki i klawiatury, a nie tylko pisząc kod.
  - Możliwość tworzenia estetycznie wyglądających tabelek – przydatna przy raportowaniu!
  - Tabele przestawne umożliwiają bardzo sprawne agregowanie danych.
  - Wygodny, jeśli chcemy stworzyć względnie proste i estetyczne wizualizacje.

Excel vs Python

- Dla wielu zastosowań, w szczególności – w przypadku relatywnie niewielkich zbiorów danych oraz braku konieczności implementowania np. metod uczenia maszynowego, Excel wystarcza.
- Python jest jednak zdecydowanie lepszy, jeśli:
  - mamy duże zbiory danych
  - mamy zamiar wielokrotnie powtarzać nasze analizy
  - mamy zamia tworzyć bardziej złożone wizualizacje
  - mamy zamiar stosować metody uczenia maszynowego lub inne bardziej skomplikowane metody statystyczne

Inne narzędzia stosowane w analizie danych

- Obok Pythona, analitycy danych często stosują język R. Jest to także oprogramowanie open-source, składnia nie różni się w istotny sposób od Pythona.
- W analizie danych stosuje się również często płatne programy, jak Excel, SPSS czy SAS.
- W pracy z bazami danych bardzo często jest używany język SQL, potrzebny do pobierania z nich danych.
- Oprogramowanie do prezentacji danych (w szczególności wizualizacji) jak np. Tableau, PowerBI.
- Oprogramowanie i metody służące do analizy bardzo dużych zbiorów danych („Big Data”): Hadoop, Spark, Microsoft Azure.
- Uwaga – analitycy danych często używają wielu narzędzi, w zależności od konkretnego celu.

Python

- Anaconda – można pobrać ją za darmo korzystając z poniższego linku:

https://repo.anaconda.com/archive/Anaconda3-2021.04-Windows-x86_64.exe

(można też skorzystać z którejś ze starszych wersji)

- Jeśli chodzi o środowisko programistyczne zachęcam do korzystania ze Spydera (dostępnego wraz z Anacondą). Można też jednak korzystać np. z edytora tekstu Atom albo Jupiter Notebook (również dostępnego z Anacondą).

3. Omówienie literatury.

Literatura i źródła on-line

- Python
  - Kursy on-line, np. na stronach www.datacamp.com albo www.edx.org.
  - Dawson M., *Python dla każdego*, Helion, 2014.
  - Bressert E., *SciPy and NumPy,* O'Reilly Media, 2012.
  - VanderPlas J., *Python Data Science Handbook*, O'Reilly Media, 2017.
    https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf
  - Dokumentacja do Pythona: https://docs.python.org/3/
  - Dokumentacja do pakietów Numpy i Scipy: http://docs.scipy.org/doc/

- Analiza danych i statystyka
  - Larose, D.T., *Odkrywanie wiedzy z danych*, Wydawnictwo Naukowe PWN, W-wa 2006.
  - Sobczyk M., *Statystyka*, Wydawnictwo Naukowe PWN, Warszawa 2002.

- Wizualizacja danych
  - Tufte E., *The Visual Display of Quantitative Information*, Graphics Press 2001.
  - *Yau N. Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics,* Wiley 2011.

www.biecek.pl/Eseje/index.html

- Inne
  - Meyer-Schonberger V., Cukier K., *BIG DATA: rewolucja, która zmieni nasze myślenie, pracę i życie*, MT Biznes, Warszawa 2014.

4. Uwagi o znaczeniu analizy danych dla współczesnej nauki, biznesu oraz życia społecznego.

Znaczenie analizy danych – uwagi wstępne

- Dane gromadzone są od powstania pisma.
  - Mezopotamia – dane dotyczący transakcji handlowych, aktów własności, demografii, itd.
  - Wszystkie imperia czy większe państwa gromadziły dane na temat różnych aspektów ich funkcjonowania.
- Operacje na danych ułatwione były najpierw dzięki rozwojowi algebry i analizy matematycznej (XIV-XVII wiek).
- Rozwój statystyki opartej na prawdopodobieństwie umożliwił stworzenie metod wnioskowania o większych populacjach na bazie losowo dobranej próby.

Znaczenie analizy danych we współczesnym świecie

- W ostatnich dziesięcioleciach lawinowo rośnie ilość wytwarzanych, oraz coraz częściej powszechnie dostępnych danych.
- Analiza danych staje się coraz bardziej istotna dla ekonomii, społeczeństwa i nauki.
- Można spotkać się z poglądem, że żyjemy w społeczeństwie informacyjnym. Pojęcie to różnie się definiuje, ale można ogólnie powiedzieć, że oznacza to, iż dystrybucja i operowanie informacją staje się istotną działalnością gospodarczą i kulturową.
- W tym kontekście pisze się czasem o „rewolucji danych”.

Ilość gromadzonych danych - przykłady

- W 2013 roku ilość przechowywanych danych na świecie wynosiła ok 1200 eksabajtów (czyli ponad bilion gigabajtów). Gdyby zapisać je w książkach, pokryłyby one powierzchnię USA 52 razy.
- Przykłady - Internet:
  - Google produkuje 24 petabajty danych każdego dnia.
  - Na Facebooku powstaje łącznie ok. 3 miliardy „lajków” i komentarzy dziennie.
- Przykład z obszaru nauki – astronomia:
  - Sloan Digital Sky Survey (przegląd nieba) rozpoczął pracę w 2000 r., w pierwszym tygodniu zebrał więcej danych niż w całej historii astronomii.
  - W 2010 roku zebranych danych było już ponad 140 terabajtów.
  - Następca – Large Synoptic Survey Telescope (Chile) zbiera tyle danych co 5 dni.

Meyer-Schonberger V., Cukier K., *BIG DATA: rewolucja, która zmieni nasze myślenie, pracę i życie*

Zjawiska, które przyczyniły się do „rewolucji danych”

- Nowe metody zbierania danych.
- Nowe metody przechowywania ich w bazach danych.
- Techniczne możliwości analizy dużych zbiorów danych.
- Spadająca cena gromadzenia i przechowywania danych.
- Powszechna dostępność narzędzi do analizy danych typu *open-source* jak Python czy R, jak również różnego rodzaju materiałów edukacyjnych dostępnych online z nimi związanych.
- Rozpowszechnienie wiedzy na temat technik analizy danych (jak np. technik uczenia maszynowego – *machine learning*).

Źródła danych (wybrane przykłady)

- Badania empiryczne
  - Ilościowe lub jakościowe dane bazujące na kwestionariuszach.
  - Dane oparte na eksperymentach – zakłada poddanie uczestników jednemu lub większej liczbie działań. Zazwyczaj mamy tu do czynienia za grupą doświadczalną i kontrolną.
- Dane dostępne z Internetu
  - Przeglądarki
  - Media społecznościowe
  - Rozmaite internetowe bazy danych
- *Internet of things*
  - Wszelkie urządzenia pobierające dane i podłączone do jakiejś sieci, dzięki której mogą się nimi wymieniać, np. do Internetu.
  - Dane pobierane np. w domach, lodówkach, czujnikach ruchu, miastach, fabrykach, knajpach, itd.

Źródła danych (wybrane przykłady cd.)

- Bazy danych związane np. z działalnością różnych organizacji (dane dotyczące transakcji, logistyki, itd., przechowywane w bazach danych firm).
- Dane demograficzne.
- Wykorzystywane przez naukę dane pochodzące z pomiarów różnych wielkości fizycznych.
- Wszelkie inne źródła danych, które przy odrobinie twórczego myślenia mogą zostać przeistoczone w przydatną informację.
  - Zakupy, aktywność w mediach społecznościowych, lokalizacja, dane dotyczące funkcji naszego organizmu itd..
  - …

(Wybrane) konsekwencje

- Duże przedsiębiorstwa, aby być konkurencyjne na dynamicznym rynku, muszą gromadzić (lub kupować) potrzebne im dane oraz inwestować w ich analizę.
- Często dostępne mamy pełną informację o danym zjawisku, tzn. o całej (albo dużej części) populacji która nas interesuje. Stąd często zdarza się, iż nie zachodzi potrzeba próbkowania danych i stosowania tradycyjnych metod statystycznych jak hipotezy statystyczne. Zamiast tego dąży się do wykrywania korelacji na masową skalę.

Wyzwania dla analizy danych

- Spójność danych pochodzących z różnych źródeł.
- Dokładność pobieranych danych.
- Odnajdywanie nowych sposobów gromadzenia i wykorzystywania danych.
- Odnajdywanie ukrytych i istotnych zależności w danych (eksploracja danych).
- Trafne i poprawne stosowanie metod predykcyjnych (w szczególności metod uczenia maszynowego).

Zagrożenia związane z analizą danych - nauka

- Nieprawidłowe korzystanie z narzędzi statystycznych.
- Niedokładność danych.
- Wybieranie danych, które pasują do z góry ustalonej tezy.
- Fałszowanie danych
  - Przykład: fałszowania danych na temat tego, że czerwone wino jest dobre na serce dopuścił się prof. Dipak Das, któremu udowodniono 145 przypadków fałszowania i fabrykowania naukowych danych, w celu wyłudzenia rządowych dotacji. Chodzi między innymi o wyniki analizy białek metodą Western blot i manipulacje dokumentującymi te badania zdjęciami.

Zagrożenia związane z analizą danych – życie społeczne

- Ograniczenie prywatności.
- Przewidywanie, kto może popełnić przestępstwo (na podstawie danych związanych z dotychczasowymi przestępstwami).
- Utrwalanie nierówności poprzez „uprzedzone modele” (np. w bankach).
- „Społeczna punktacja” oparta na danych w Chinach – nadużywanie danych?

5. Podstawowe pojęcia związane z analizą danych.

Podstawowe pojęcia związane z analizą danych

- **Przetwarzanie danych**
  - Zapis, odczyt
  - Proste przekształcenia
  - Przygotowywanie danych do dalszych analiz

- **Analiza danych**
  - Przetwarzanie danych
  - Analizy z zakresu statystyki opisowej (opis statystyczny zmiennej)
  - Klasyczne metody jak hipotezy statystyczne
  - Łączenie danych z różnych źródeł, uspójnianie danych

Podstawowe pojęcia związane z analizą danych

- **Data mining** lub **eksploracyjna analiza danych** (*data driven discovery, exploratory data analysis*)
  - Odnajdywanie prawidłowości czy regularności w danych oraz zależności pomiędzy zmiennymi itd.
  - Wydobywanie informacji, która jest „ukryta” w „gąszczu danych”, tzn. jest w praktyce niemożliwa do zauważenia bez zastosowania pewnych metod.

- **Wizualizacja danych**
  - Graficzna prezentacja danych – zastosowanie różnych typów wizualizacji zawierających takie graficzne elementy jak krzywe, słupki, kolory, itd. w celu przekazania wybranych zależności zawartych w danych (zazwyczaj w celu ich lepszego zrozumienia, wykrycia zależności za pomocą zmysłu wzroku).
  - Stosowana zazwyczaj w celu eksploracji bądź raportowania danych.

Podstawowe pojęcia związane z analizą danych

- **Uczenie maszynowe (*machine learning*)**
  - Dziedzina zajmująca się tworzeniem programów (poprzez konstruowanie algorytmów i stosowanie metod statystycznych), które się *uczą*. Przez uczenie się zazwyczaj rozumiemy tutaj automatyczne poprawianie skuteczności algorytmu (w zakresie np. predykcji) wraz z dostarczaniem nowych danych. Algorytmy poprawiają swoje parametry, czy raczej dostosowują je do nowych danych.
  - Metody te są stosowane w celu eksploracji danych oraz m.in. predykcji

- **Data science**
  - „Klasyczna” analiza danych poszerzona o zastosowania metod i algorytmów uczenia maszynowego.
  - „Data scientist” powinien posiadać szeroką wiedzę zarówno ze statystyki, metod uczenia maszynowego jak i techniczne umiejętności związane z programowaniem czy pracą z bazami danych.

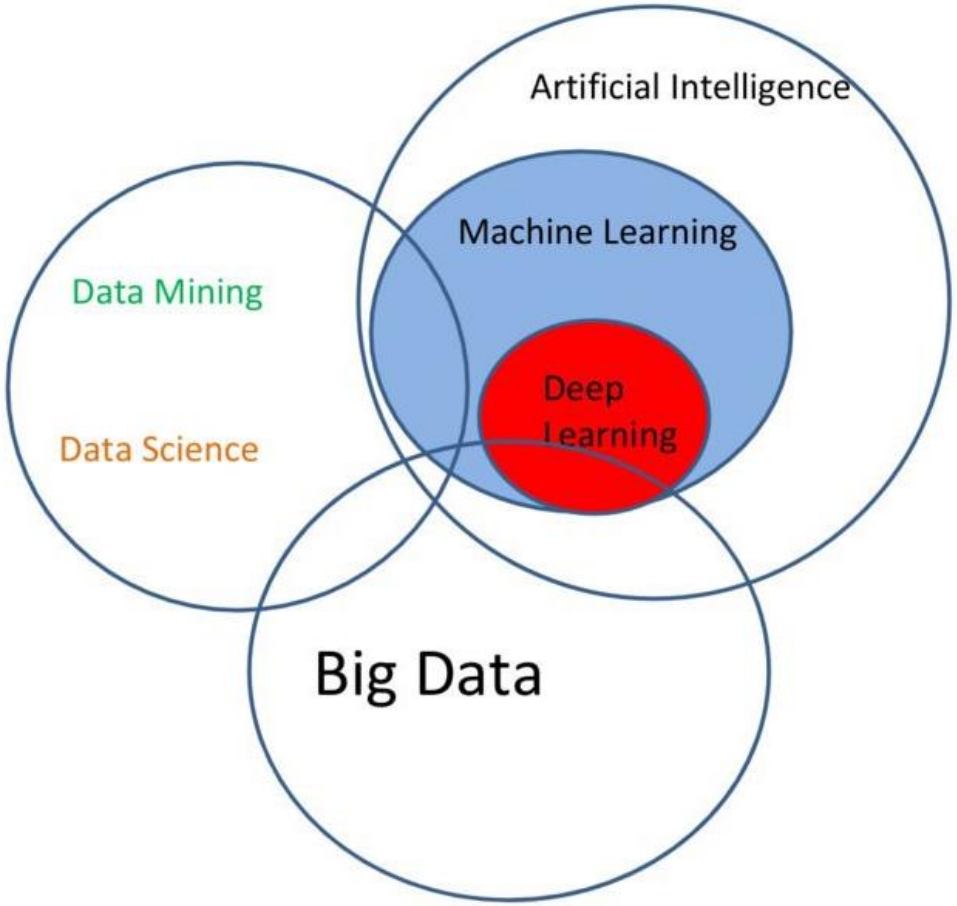
Podstawowe pojęcia związane z analizą danych

- **Sztuczna inteligencja** (*Artificial Intelligence* – AI)
  - Szeroki wachlarz metod, które mają służyć naśladowaniu zachowania lub procesów kognitywnych człowieka bądź automatyzowaniu jakiejś czynności przez niego wykonywanej. Metody uczenia maszynowego mogą służyć takim celom, stąd często przyjmuje się, że należą do dziedziny AI.

- **Data engineering**
  - Projektowanie baz danych oraz administracja bazami danych, jak równiez dbanie o poprawność ich funkcjonowania, np. o prawidłowy przepływ danych.

- **Big data**
  - Ogólny termin odnoszony do samej wielkości zbiorów danych (zbiory tak duże, że przetwarzanie ich za pomocą komputerów osobistych nie jest możliwe) albo też do pewnego zjawiska społecznego.

Jeden z możliwych sposobów rozumienia zależności pomiędzy omawianymi pojęciami



Co mówi nam Google Trends o zainteresowaniu analizą danych

Big data
Wyszukiwane hasło

Machine learning
Wyszukiwane hasło

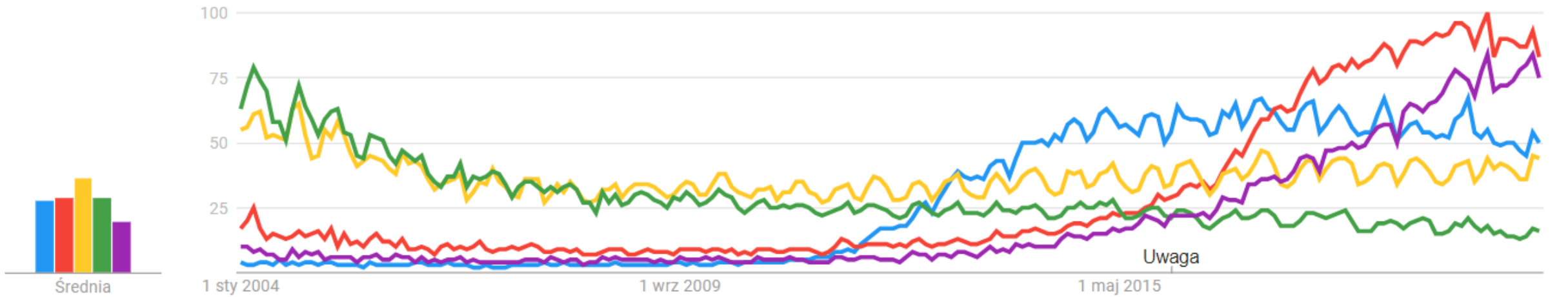
Data analysis
Wyszukiwane hasło

Data mining
Wyszukiwane hasło

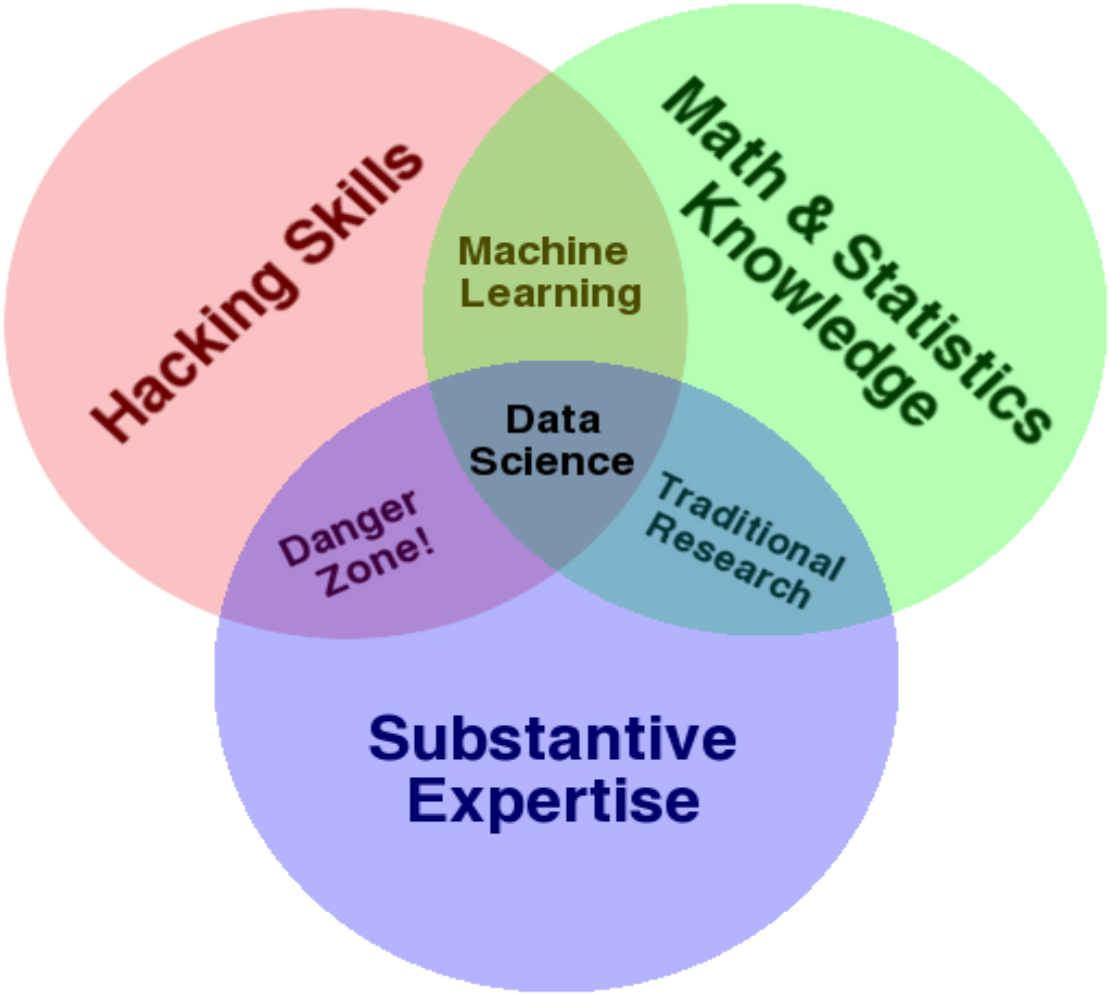
Data science
Wyszukiwane hasło

Cały świat | 2004 – dziś | Wszystko | Wyszukiwarka Google

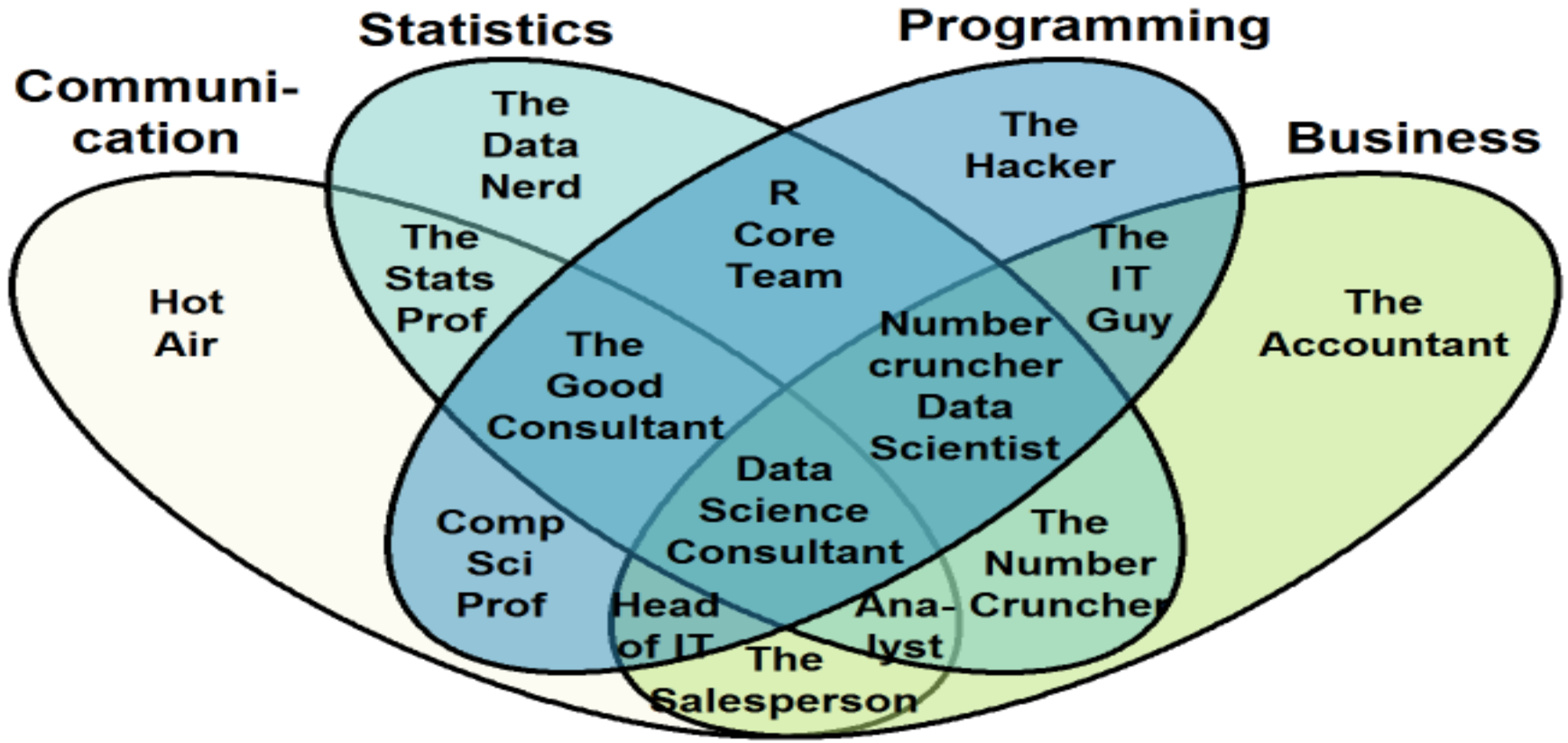
Zainteresowanie w ujęciu czasowym



Data Science Venn Diagram – jakie umiejętności powinien posiadać „data scientist”



The Data Scientist Venn Diagram



Dane a informacja

Pojęcie informacji trudno jest zdefiniować, stąd w literaturze przedmiotu istnieje wiele propozycji. Podobnie, różnie można rozumieć różnicę pomiędzy danymi a informacją. Przyjmiemy tutaj następujące rozumienie tych pojęć:

- Dane mają charakter materialny – składają się z numerów, napisów, itd. Można powiedzieć, że reprezentują fakty, będąc kodowane za pomocą odpowiednich symboli w taki sposób, aby mogły być łatwo zapisywane, przetwarzane i przesyłane.
- Informacja to zinterpretowane dane – tak, aby mogły pogłębić nasze rozumienie danego zjawiska albo ułatwić podjęcie jakiejś decyzji.
  - Uwaga – aby dane można było w ten sposób zinterpretować, należy zazwyczaj je odpowiednio przekształcić i poddać analizie np. statystycznej.

Dane a informacja

- Informacja ma charakter subiektywny – z tych samych danych różni ludzie mogą „wyciągać” różne informacje.
  - To, jaka informacja będzie „wyciągnięta” z konkretnego zbioru danych zależy często od specyficznego celu analiz, ich odbiorcy oraz aktualnie posiadanej wiedzy.
- Często wymienia tu się również trzeci element – *wiedzę*, która powstaje z informacji.
- **Jednym z podstawowych zadań analizy danych jest przekształcenie danych w informację.** Zadanie jest to szczególnie ważne współcześnie, z uwagi na ogromną ilość dostępnych danych, których użyteczność i znaczenie trzeba „wydobyć” za pomocą analiz i wizualizacji.

6. Uwagi o strukturze danych z którymi będziemy mieli do czynienia i formatach plików z danymi.

Wstępne uwagi o strukturach danych

- Istnieje wiele typów struktur danych i ich modeli. Nie jest naszym celem omawianie ich wszystkich. Opiszemy tu jedynie wstępnie typ struktury danych, którym będziemy się zajmować na naszych zajęciach. Jest to bardzo często stosowany model danych, opiera się na nim duża część metod statystycznych, metod analizy danych czy algorytmów uczenia maszynowego.
- Dane, które będziemy wykorzystywać to dane tabelaryczne (w przeciwieństwie np. do danych hierarchicznych) szczególnego typu.

Wstępne uwagi o strukturach danych

. W tym modelu zakładamy, że:

1) Każdy rząd odpowiada obserwacji (przypadkowi) z badanej populacji (może być to np. respondent w badaniu).

2) Kolumnom odpowiadają intuicyjnie własności, cechy naszych obserwacji. Cechy te określa się również jako *zmienne*, której nazwy będziemy najczęściej używać. W języku angielskim używa się np. terminów *variable* czy *attribute;* w przypadku zastosowań algorytmów uczenia maszynowego używa się też nazw *predictors* lub *features*.

3) W każdej kolumnie zawarte są dane jednego typu.

4) Każda kolumna odpowiada dokładnie jednej zmiennej (cesze). W jednej kolumnie nie może być zawartych kilku zmiennych, a jedna zmienna nie może być „rozproszona” pomiędzy większą liczbę kolumn.

5) Kolumny mają zazwyczaj etykiety, rzędy nie muszą mieć etykiet, ale mogą.

5) W jednej tabeli przechowywany jest tylko jeden typ jednostki obserwacyjnej.

6) Podobnie, każdej obserwacji odpowiada dokładnie jeden wiersz. W danych powinna być zawarta zmienna identyfikująca obserwacje.

Czasem mówi się, że dane spełniające wszystkie wyżej wymienione warunki są *tidy* (dobrze uporządkowane).

Sprowadzenie danych do takiego formatu jest zazwyczaj wysoce pożądane – umożliwia poprawne stosowanie różnorodnych metod analitycznych i statystycznych.

Uwaga: czasem to, co potraktujemy jako „obserwację” zależy od kontekstu i celu analiz.

Przykład

ID	Imię	Nazwisko	Wiek	Nr. telefonu	Funkcja	Nr. pokoju
01	Marcin	Kowalski	33	100-200-300	Programista	14
02	Edyta	Schmidt	29	201-202-203	Analityk danych	18
03	Anna	Wiśniewska	41	900-900-900	Sprzedawca	22

Dane nieuporządkowane - przykład

year	month	element	day1	day2	day3	day4	day5
2014	1	temp_max	-	3	8	-	3
2014	1	temp_min	-	1	4,4	-	0
2014	2	temp_max	5	-	2	-	-
2014	2	temp_min	1	-	-4	-	-

Dane nieuporządkowane - przykład

year	month	element	day1	day2	day3	day4	day5
2014	1	temp_max	-	3	8	-	3
2014	1	temp_min	-	1	4,4	-	0
2014	2	temp_max	5	-	2	-	-
2014	2	temp_min	1	-	-4	-	-

date	element	temperature
02.01.2014	temp_max	3
02.01.2014	temp_min	1
03.01.2014	temp_max	8
03.01.2014	temp_min	4,4
05.01.2014	temp_max	3
05.01.2014	temp_min	0
01.02.2014	temp_max	5
01.02.2014	temp_min	1
03.02.2014	temp_max	2
03.02.2014	temp_min	-4

Dane nieuporządkowane - przykład

year	month	element	day1	day2	day3	day4	day5
2014	1	temp_max	-	3	8	-	3
2014	1	temp_min	-	1	4,4	-	0
2014	2	temp_max	5	-	2	-	-
2014	2	temp_min	1	-	-4	-	-

date	element	temperature
02.01.2014	temp_max	3
02.01.2014	temp_min	1
03.01.2014	temp_max	8
03.01.2014	temp_min	4,4
05.01.2014	temp_max	3
05.01.2014	temp_min	0
01.02.2014	temp_max	5
01.02.2014	temp_min	1
03.02.2014	temp_max	2
03.02.2014	temp_min	-4

date	temp_max	temp_min
02.01.2014	3	1
03.01.2014	8	4,4
05.01.2014	3	0
01.02.2014	5	1
03.02.2014	2	-4

Tidy data!

Struktura danych - uwagi

- Dane miewają bardzo różne formaty, w zależności od ich źródła, sposobu ich zbierania i narzędzia za pomocą którego są zakodowane.
- Dodatkowo zdarza się, że dane (w szczególności np. te ściągnięte z Internetu, ale także inne rodzaje danych) są nieuporządkowane i pełne błędów. Stąd zachodzi potrzeba czyszczenia danych oraz sprowadzenia ich (o ile to możliwe) do formatu, który opisaliśmy.
- Istnieje wiele różnych możliwych sposobów, na które dane mogą być nieuporządkowane oraz wiele możliwych problemów z danymi. Ważniejsze z nich to np. braki danych oraz wartości odstające – omówimy je na kolejnym wykładzie.

Typy zmiennych (powtórka ze statystyki)

Wspomniane powyżej cechy obserwacji są różnego typu. Można je najogólniej podzielić na numeryczne i kategorialne, te z kolei na dalsze podgrupy w zależności od skali, na jakiej są mierzone.

1) Skala ilorazowa (stosunkowa)
2) Skala interwałowa (przedziałowa)

ilościowe (numeryczne)

3) Skala porządkowa (rangowa)
4) Skala nominalna

jakościowe (kategorialne)

Czym różnią się od siebie dwie skale jakościowe?

- Skala nominalna
  - Nie można w sensowny sposób mówić o porządku pomiędzy kategoriami danych ani o operacjach arytmetycznych na nich.
  - Przykłady: płeć, opinie polityczne, miejsce urodzenia.

- Skala porządkowa
  - Skala jest porządkowa, kiedy kategorie można w jakiś sposób uporządkować od najmniejszej do największej.
  - Nie można jednak w ich przypadku w sensowy sposób wykonywać operacji arytmetycznych, ponieważ „odległość” pomiędzy kategoriami nie jest jasno określona.
  - Przykłady: miejsce na podium, ocena.

Czym różnią się od siebie dwie skale ilościowe?

- Skala interwałowa
  - Nie istnieje wyróżniony element zerowy – jest umowny i nie oznacza „braku” pewnej cechy.
  - Dozwolone są tylko niektóre operacje – dodawanie i odejmowanie.
  - Odległość ma konkretne znaczenie.
  - Przykłady: temperatura w stopniach Celsjusza, data kalendarzowa.
  - Iloraz nie ma fizycznego znaczenia (nie możemy np. powiedzieć, że 01.01.1000 to dwa razy później niż 01.01.2000).

- Skala ilorazowa:
  - Istnieje wyróżniony element zerowy, który ma interpretację fizyczną – zazwyczaj „brak” danej cechy.
  - Dozwolone są wszystkie operacje.
  - Przykłady: temperatura w Kelvinach, wysokość, ciężar.
  - Można w tym przypadku sensownie powiedzieć: „x jest 2 razy wyższy niż y”.

Typy danych w Pythonie

- Podstawowe typy danych w Pythonie:

➢ Dane numeryczne: **int**, **float**, **complex**

➢ Dane tekstowe: **str**

➢ Dane logiczne: **bool**

➢ Listy: **list**

Format pliku „.csv”

- CSV to skrót od *Comma Separated Values*
- Dane tabelaryczne przechowywane jako tekst.
- Poszczególne pola w każdym wierszu odpowiadają kolumnom. Rozdzielone są separatorem, którym najczęściej jest przecinek (czasem średnik).
- Wartości pól mogą być ujęte w cudzysłowy.
- Wartości zawierające używany znak separatora muszą być ujęte w cudzysłowy.
- Pierwsza linia może stanowić nagłówek zawierający nazwy pól rekordów, jednak pierwszy wiersz pliku CSV wg standardu ma takie samo znaczenie jak pozostałe.

Format pliku „.csv”

imię, zawód, wiek
Andrzej Nowak,kierowca,42
"Jan Wiśniewski",fryzjer,24
Janina Kowalska,wykładowca,50

imię	zawód	wiek
Andrzej Nowak	kierowca	42
Jan Wiśniewski	fryzjer	24
Janina Kowalska	wykładow	50

Format pliku „.txt”

- W pliku o rozszerzeniu .txt może być z zasady zawarty dowolny tekst.
- Jeśli używamy takiego pliku do przechowywania danych tabelarycznych, możemy rozdzielić wartości należące do kolejnych kolumn dowolnym znakiem (*separatorem*). Najczęściej używa się tu tabulacji spacji oraz znaków „,”, „;”.
- Pliki tekstowe z dowolnymi separatorami można z łatwością odczytać zarówno w Pythonie jak i w Excelu.

Inne formaty plików

- Pliki „Excelowe” (.xlsx)
- Pliki natywne dla SPSS (.dat)
- Pliki json (.json)
- Pliki .xml
- Inne..