1 Tematy projektów do kursu IM148

Każdy uczestnik kursu IM148 wybiera jeden temat projektu i samodzielnie go realizuje. Pod koniec kwietnia podczas zajęć będzie rozesłany email z linkiem do arkusza, gdzie się będzie można wpisywać kto co robi. Każdy student w grupie powinien robić inny temat.

Projekty są robione w okresie 2–3 ostatnich zajęć. Przynajmniej jeden raz trzeba **skonsultować** swoje postępy z realizacji projektu lub omówić działanie swojego programu **podczas tych zajęć**. Przy 12 osobach w grupie przypada ok. 20 minut na osobę. **Bez konsultacji można dostać maksymalnie połowę punktów za projekt.** W arkuszu wyboru projektu będzie również lista z zapisami na terminy tych konsultacji, żeby nie zostało za dużo na sam koniec zajęć.

Skończony program proszę zamieścić jako odpowiedź do zadania PROJEKT na pegaz.uj.edu.pl, do końca semestru letniego (wpis w pierwszym terminie) lub do końca sesji poprawkowej (wpis w drugim terminie).

Prawie wszystkie projekty da się zrobić bez pomocy specjalistycznych bibliotek. Najbardziej zaawansowanym pakietem potrzebnym w połowie przypadków jest `matplotlib` Nie należy korzystać z modułu `pandas`, bo nie nadaje się dla początkujących.

1.1 masa molowa

(analiza stringów, chemia)
Napisz moduł `molarmass` z funkcją `mm()` która wylicza masę molową podanego związku. Funkcja powinna działać dla pierwiastków od H do Bi tak, żeby dało się jej użyć jak poniżej:

```
from molarmass import mm
print( mm('C2 H5 O H') )
print( mm('C8H10N4O2') )
print( mm('Na0.75K0.25Cl1I0.002') )
```

1.2 ile wody płynie w Wiśle

(operacje na plikach, analiza danych)
Pod adresem https://dane.imgw.pl/data/dane_pomiarowo_obserwacyjne/dane_hydrologiczne/dobowe/ są dane z wodowskazów w całej Polsce z ostatnich 50 lat. Napisz program, który rysuje poziom wody w Wiśle (wodowskaz Kraków-Bielany) w latach hydrologicznych 2016–2018. Kiedy były powodzie? Potrzebne będzie: ręczne ściągnięcie 36 plików zip, rozpakowanie jednego i rozpoznanie jakie dane w formacie csv są tam zapisywane. Program powinien w locie rozpakowywać te pliki, wczytywać potrzebne dane i robić odpowiedni wykres. Ponadto program powinien wyliczyć ile wody przepłynęło przez Kraków w roku 2017 (dane dla Kraków-Czernichów zawierają takie dane).

1.3 genom wirusa SARS-CoV-2

(bio, wczytywanie i analiza danych)
Genom wirusa SARS-CoV-2 został opublikowany i jest dostępny w bazie https://www.ncbi.nlm.nih.gov/nuccore/MN908947 Więcej szczegółów w artykule z 19 lutego 2020 *Coronavirus and the race to distribute reliable diagnostics* https://www.nature.com/articles/d41587-020-00002-2 Inny artykuł https://academic.oup.com/clinchem/advance-article/doi/10.1093/clinchem/hvaa029/5719336 podaje krótkie sekwencje genów, których można użyć do diagnostyki metodą PCR. Są to: *primers (forward and reverse)* oraz *probes*.

Napisz program, który szuka tych sekwencji w genomie SARS-CoV-2. Może to wymagać odpowiednich modyfikacji tych sekwencji (na komplementarne, albo odwrócone). Program nie powinien używać specjalistycznych bibliotek i robić wszystko przy pomocy standardowych funkcji.

1.4 prezentacja błądzenia losowego

(symulacje, wykresy)
Błądzenie losowe (*random walk)* to model opisujący ewolucję jakiejś zmiennej na skutek pojedynczych losowych kroków tej zmiennej. Może być użyty do opisu niektórych zjawisk w finansach, chemii, biologii, czy fizyce polimerów. Więcej szczegółów: https://en.wikipedia.org/wiki/Random_walk lub polski odpowiednik. Wykonaj symulację obrazującą błądzenie losowe na dwuwymiarowej sieci kwadratowej. Wykonaj 2 wykresy jak najbardziej podobne do tych na powyższych stronach (Wykresy ruchu ośmiu cząstek otrzymane w symulacji błądzenia losowego.), (Random walk in two dimensions with 25 thousands steps).

1.5 wynik błądzenia losowego

(symulacje, wykres)
Stu pijaków chodzi po kartce w kratkę. Wszyscy startują z pola $(0, 0)$ i robią losowo krok na sąsiednie pole. Jak daleko zajdą po n krokach? Przeprowadź taka symulację i wyznacz jaka jest średnia (kwadratowa) odległość d po n krokach. Narysuj wykres d w funkcji n, razem z krzywą $\sqrt{n}$, w zakresie do $n = 1000$.

1.6 prey-predator model

(symulacje, równania różniczkowe, wykresy)
Liczba futer zajęcy i rysi sprzedanych przez Hudson's Bay Company w latach 1845–1935 wykazuje wyraźnie oscylującą zależność od czasu (patrz pierwszy wykres na stronie https://en.wikipedia.org/wiki/Lotka-Volterra_equations) Takie zależności wynikają ze zmian liczebności ofiar i drapieżników i można to próbować symulować stosując podane na tej stronie równania różniczkowe Lotka–Volterra.

Napisz program robiący taką symulację i zrób w nim dwa wykresy, jak najbardziej zbliżone do dwu wykresów po prawej na powyższej stronie wikipedii. (Population dynamics for baboons and cheetahs problem mentioned aside.) oraz (Phase-space plot for the predator prey problem for various initial conditions of the predator population.)

Podobne symulacje i artykuł na ten temat: https://www.digitalbiologist.com/blog/2018/9/a-population-dynamics-model-in-five-lines-of-python

1.7 wielomiany Czebyszewa

(matematyka, informatyka)
Znajdź na wikipedia.org co to są wielomiany Czebyszewa T_k. Wykorzystaj rekurencyjną definicję tych wielomianów, żeby wyliczyć współczynniki dowolnego wielomianu T_k i stwórz tablicę tych współczynników dla $k = 1, \ldots, 20$. Wypisz te wielomiany w czytelnej postaci i porównaj z podanymi na wikipedii kilkoma pierwszymi, żeby sprawdzić czy program dobrze działa. Użyj otrzymanych wielomianów do zrobienia wykresów kilku $T_k(x)$ w przedziale $(-1.1, 1.1)$.

1.8 klasa wielomianów

(matematyka, programowanie obiektowe)

Napisz funkcję, która wylicza wielomian o podanych współczynnikach, dla podanego argumentu. Zastanów się jak prosto reprezentować wielomian. Nastepnie napisz klasę, która działa na wielomianach tak, żeby można było:

```
w1 = Wielo([1,0,3])         # stworzyć wielomian np. 1 + 3*x**2
w2 = Wielo('1+3x^2')        # to ambitniejsza wersja, niekoniecznie potrzebna
w1.val(x)                   # wyliczać wartość dla podanego x
w1.stopien                  # dostać stopień wielomianu
w4 = w1+w2                  # dodawać wielomiany
w5 = w1*w2                  # mnożyć wielomiany
w6 = 4*w4+w5                # mnożyć przez liczbę
str(w6)                     # zwraca string z wielomianem w czytelnej formie, np. '1 + 3x^2'
```

1.9 Wordle

(łamigłówki, algorytmy)
Na stronach The New York Times jest gra Wordle, która zrobiła furorę w 2022. https://pl.wikipedia.org/wiki/Wordle Napisz program, który w jakimkolwiek stopniu pomaga zdobyć dobry rezultat w tej grze, lub wybrać dobrą strategię. Na jeden dzień jest tylko jedno słowo do zgadnięcia, ale można znaleźć w sieci archiwalne zagadki, gdzie można testować swój program.

1.10 funkcja logistyczna i chaos

(matematyka, symulacje)
Przedstaw na wykresie wartości 1000 kolejnych x_n dla iteracyjnie wyliczanej funkcji logistycznej, tzn. $x_{n+1} = a * x_n * (1 - x_n)$, dla $a = 3.9$, dla jakiegoś wybranego x_0 z przedziału (0,1). Zrób próby jak to zachowanie wygląda dla kilku innych wartości a. Zrób wykres złożony z punktów (a, x_n), na którym będą naraz wszystkie wartości dla $n = 100 \ldots 10000$, dla a z przedziału $(1, 4)$.

Dalsze informacje: Film: *This equation will change how you see the world (the logistic map)*
https://www.youtube.com/watch?v=ovJcsL7vyrk
https://en.wikipedia.org/wiki/Logistic_map
https://upload.wikimedia.org/wikipedia/commons/0/0a/Subsection_Bifurcation_Diagram_Logistic_Map.png

1.11 seriale

(operacje na plikach, analiza danych, wykres)
Zrób wykres przedstawiający średnią ocenę widzów dla poszczególnych odcinków ulubionego serialu. Można zobaczyć tego typu wykresy https://i.redd.it/h3yxb4tmvgp41.png opublikowane na forum https://www.reddit.com/r/dataisbeautiful/

Średnie oceny różnych filmów są podane w Internet Movie Database imdb.com Wyniki można ściągnąć w formacie tekstowym ze strony https://datasets.imdbws.com/ Są to duże pliki tekstowe, krótki opis formatu jest na http://www.imdb.com/interfaces/

Należy ręcznie sciągnąć te pliki na swój komputer i rozpakować (być może trzeba pozmieniać nazwy po rozpakowaniu). Dalej, przy pomocy skryptu w Pythonie należy:

- w pliku `title.basics.tsv` znaleźć kod danego serialu,
- w pliku `title.episode.tsv` znaleźć kody poszczególnych odcinków,
- w pliku `title.ratings.tsv` znaleźć oceny poszczególnych odcinków,

- te dane warto zebrać tymczasowo w małą tablicę i zapisać do pliku do dalszych szybkich testów,
- zrobić z tych danych wykres.

Na koniec trzeba zrobić taką funkcję, która dostaje jako parametr tytuł serialu i robi automatycznie wykres, np. `plot_ratings('Breaking Bad')`

1.12 wybuchy na Słońcu

(astronomia, analiza danych, wykres)
Sonda kosmiczna SOHO obserwuje Słońce. Jest w niej instrument LASCO, który obserwuje koronę słoneczną i zbiera informacje o Coronal Mass Ejections (wybuchach plazmy wyrzucanej ze Słońca). Więcej informacji: https://en.wikipedia.org/wiki/Solar_and_Heliospheric_Observatory
https://en.wikipedia.org/wiki/Large_Angle_and_Spectrometric_Coronagraph

Takie zdarzenia są zebrane w pliku tekstowym https://cdaw.gsfc.nasa.gov/CME_list/UNIVERSAL/text_ver/univ_all.txt (Do wstępnych testów warto sobie ten plik obciąć do kilku linii.) Napisz program, który wczytuje daty i energię kinetyczną wybuchów z tego pliku i rysuje czasowy histogram tych zdarzeń (ile w danym roku) biorąc pod uwagę tylko te, dla których energia kinetyczna `>1e+30` Czy widać w tym cykl aktywności Słońca?

1.13 gwiazdy

(astronomia, analiza danych, wykres)
Zrób wykres absolutnej jasności w funkcji wskaźnika barwy dla dużego zbioru gwiazd, na podstawie danych https://github.com/astronexus/HYG-Database Jest tam plik tekstowy `hygdata_v3.csv` (32 MB) oraz opis formatu. Potrzebne są tylko kolumny 'color index' oraz 'abs mag'. Do testów warto sobie ten plik obciąć do kilku linii.

Przykładowy projekt pod Pythonem, który robi coś podobnego https://github.com/RobertoIA/Hertzsprung-Russell Można się na tym wzorować, ale proszę to uprościć jak najbardziej. Plik ściągnąć ręcznie na lokalny dysk. Bez bibliotek panda. Bez kolorów. Bez animacji. Opisane osie, dobrane skale.

Czy widać ciąg główny? Czy widać karły i olbrzymy? Więcej informacji na stronie https://en.wikipedia.org/wiki/Hertzsprung%E2%80%93Russell_diagram

1.14 jednoosobowe scrabble

(gry i ich zasady, struktury danych, operacje na plikach)
Trzeba napisać grywalne, jednoosobowe scrabble. Komputer losuje 7 liter z zestawu i trzeba ułożyć z tych liter słowo po angielsku. Komputer sprawdza czy to słowo jest na liście dozwolonych słów i liczy punktację wg. reguł Scrabble. Trzeba zaproponować jak traktować złe odpowiedzi i to uwzględnić w programie, ile ma być pytań, i ewentualnie wprowadzić inne reguły, tak żeby gra była grywalna.

1.15 Hall of Fame

(operacje na plikach)
Napisz moduł do obsługi listy najlepszych wyników uzyskanych w grze. Lista powinna być pamiętana w pliku tekstowym i zawierać 5 najlepszych wyników wraz z imieniem gracza, podawanym przez niego gdy ma być wpisany na listę. Powinna zawierać funkcje sprawdzenia czy dany wynik

gry wystarcza do wpisania na listę, uaktualnienia listy o nowy wynik, ładnego drukowania listy i ewentualnie odgrywania fanfarów dla zwycięzcy.

Przetestuj użycie tego modułu w prymitywnej grze, która polega na losowaniu jednej liczby jako wynik gry.

1.16 kółko i krzyżyk

(symulacje, struktury danych i algorytmy)
Panowie O. i X. grają w kółko i krzyżyk na planszy 3×3 całkowicie losowo wybierając pola. Przeprowadź symulację 100000 rozgrywek i podaj średnie wyniki (wygrane O, wygrane X, remisy), żeby zobaczyć na ile pomaga rozpoczynanie w tej grze. Dodatkowo sprawdź na ile pomaga lub przeszkadza w wygranej zajęcie konkretnego pola w pierwszym ruchu (są tylko 3 możliwości).

Plansza w programie powinna być reprezentowana przez jednowymiarową listę liczb (koniecznie!).

1.17 maraton

(statystyka, wykres, operacje na plikach)
Na stronie `maratonwarszawski.com` są pełne wyniki archiwalne tej imprezy, np. z 2019 roku jako plik xlsx (4500 uczestników którzy ukończyli bieg z czasami od 2:11:39 do 6:44:24).

Zrób na podstawie tych danych wykres pudełkowy, porównując wyniki mężczyzn i kobiet. Zrób podobny wykres porównujący kategorie M20, M30, M40, M50, M60 i M70.

Tutorial jak robić takie wykresy w matplotlib jest na `https://ichi.pro/pl/wykres-pudelkowy-w-pythonie-kompletny-przewodnik-dla-poczatkujacych-171833219901815`
Można pod Excelem wyeksportować te dane lub ich część do pliku csv i taki plik wczytywać w swoim skrypcie. Proszę nie używać bibliotek `pandas` ani `seaborn`. Wystarczą do tego moduły `matplotlib.pyplot`, i ewentualnie `csv`

Pomysł zadania pochodzi z bardzo ładnej książeczki *Wykresy unplugged* `https://betaandbit.github.io/WykresyUnplugged/`

1.18 AA

(operacje na plikach, struktury danych, algorytmy dopasowania tesktu)
Napisz program wyliczający zawartość etanolu w drinkach. Coś podobnego jest na stronie National Institute of Health. `https://www.rethinkingdrinking.niaaa.nih.gov/Tools/Calculators/Cocktail-Calculator.aspx` Program powinien mieć dane wejściowe w 2 plikach tekstowych o czytelnym formacie. W jednym zawartość procentową etanolu w różnych składnikach. W drugim: przepisy różnych drinków (składniki oraz ilość). Program powinien używać jednostek układu SI oraz przeliczać to na 1 piwo.

1.19 rozpad radioaktywnego ^{14}C

(symulacje, wykres)
Wykonaj 10 symulacji, każda trwająca 15000 lat (rok po roku), w której $N_0 = 50$ jąder ^{14}C rozpada się. Prawdopodobieństwo, że dane jądro rozpadnie się w ciągu 1 roku wynosi $p = 0.000121$/rok. W symulacji, jeżeli `random.random()<p` to jądro się rozpada. Pokaż liczbę jąder ^{14}C w funkcji czasu, dla wszystkich 10 symulacji naraz. Narysuj na tym tle teoretyczną krzywą $N(t) = N_0 \exp(-pt)$.

1.20 PM10

(analiza danych, wykres, trudniejsze)
Na stronie https://powietrze.gios.gov.pl/pjp/archives są zebrane wyniki archiwalnych pomiarów zanieczyszczeń powietrza. Ściągnij całe dane z 2019 roku (plik zip), wypakuj z tego ręcznie plik `2019_PM10_1g.xlsx`, wytnij z niego kolumny tak by została tylko jedna wybrana stacja w Krakowie (stacje 42-49) i zapisz ten plik w formacie csv.

Następnie napisz program w Pythonie, który:

1. Wczytuje te dane. Trzeba jakoś sobie poradzić z brakującymi danymi.
2. Uśrednia te dane (też trzeba jakoś omijać brakujące pomiary) na średnie dobowe. Ile razy została przekroczona dopuszczalna norma (trzeba poszukać jaka jest i ile razy może zostać przekroczona w ciągu roku).
3. Uśrednia te dane w inny sposób, tak żeby zobaczyć jakie są średnie dobowe zmiany (w jakich godzinach są największe zanieczyszczenia – w szczycie komunikacyjnym czy w godzinach kiedy palą w piecach).
4. Zrób dwa wykresy prezentujące te obserwacje.

1.21 PM2.5

(wczytywanie danych, analiza danych)
Na podstawie danych jak powyżej trzeba zweryfikować listę rekordzistów https://www.iqair.com/world-most-polluted-cities?continent=59af92ac3e70001c1bd78e52&country=7bvZck9Ky2CAQaCiH&state=&page=1&perPage=50&cities= Żadnych wykresów nie potrzeba, przez co projekt jest prostszy niż poprzedni. Może być rok 2019. Jakie nadające się do mieszkania miejsca w Polsce mają czyste powietrze?

1.22 kiedyś to były zimy

(wczytywanie i analiza danych, wykres)
Sprawdź kiedy była najmroźniejsza zima w latach 2001-2020. Minimalne, maksymalne i średnie dobowe temperatury są dostępne w archiwum https://danepubliczne.imgw.pl/data/dane_pomiarowo_obserwacyjne/dane_meteorologiczne/ Można to zrobić dla stacji meteorologicznej Kraków-Obserwatorium, albo innej wybranej. Na potrzeby tego projektu wystarczy sprawdzić styczeń każdego roku.

Przykładowo potrzebny jest plik https://danepubliczne.imgw.pl/data/dane_pomiarowo_obserwacyjne/dane_meteorologiczne/dobowe/klimat/2010/2010_01_k.zip
Opis formatu danych jest w nadrzędnym katalogu.

Ustal jakiś sposób, jak decydować która zima była najmroźniejsza. Zrób jakiś wykres na potwierdzenie swojego wniosku.

1.23 deszcz

(wczytywanie i analiza danych)
Na podstawie danych jak w poprzednim projekcie, sprawdź w których stacjach meteo w Polsce spadło z najwięcej deszczu. Wypisz posortowane sumaryczne dane za 2020 rok.

1.24 trojaczki

(algorytmy, struktury danych, złożoność obliczeniowa)
Stwórz listę z N losowymi liczbami całkowitymi z przedziału $[1, N]$ dla jakiegoś ustalonego N. Wymyśl i zaimplementuj dwa całkiem różne algorytmy, które znajdują trójki równych liczb na tej liście (bez żadnych powtórzeń, tzn. jeżeli są cztery liczby równe 7, to jest to liczone jako jedna trójka). Upewnij się, że te algorytmy dają identyczne wyniki.

Zmierz czas potrzebny na wykonanie takiego szukania w zależności od N dla tych algorytmów. Jak koszt wykonania rośnie z N? $O(N)$, $O(N \log N)$, $O(N^2)$, $O(N^3)$? Czy da się zrealizować algorytm o koszcie $O(N)$? Proszę nie importować niczego poza `import random, time`

1.25 zdania

(lingwistyka kognitywna, proste)
Testy wykazują, że jeżeli w zdaniu w każdym wyrazie pozmieniamy kolejność liter, zachowując pierwszą i ostatnią na swoich miejscach, to można przeczytać całe zdanie bez problemu.

Napisz program, który dla wczytanego z pliku tekstu złożonego z kilku zdań, przerabia go losowo według takiej zasady i sprawdź czy łatwo czytać takie zdania. Sprawdź to dla języka polskiego i angielskiego.

Npasiz pagrorm, króty dla wngzeaytco z pkliu ttesku zżoogneło z kkilu zadń, pazibrrea go lsowoo wdłeug tkieaj zaasdy i srpdwaź czy łtawo cytazć tkaie zdinaa. Sdarpwź to dla jękzya pgeksiolo i aenelgkigsio.

1.26 projekty na inne tematy

Można zaproponować inny temat. Jest to możliwe tylko do 15tego maja, a później już nie. Należy napisać maila do prowadzącego precyzując w kilku linijkach co będzie robił program. Nie może to być żaden oklepany temat, na który można znaleźć różne gotowe programy. Należy się spodziewać, że prowadzący pracownię doda do specyfikacji projektu jakieś wymagania.