Языки обработки данных

Судаков Владимир Анатольевич

sudakov@ws-dss.com

2019

План

- Введение/Знакомство
- Python
  - Динамические аспекты и ООП
  - Объектная модель
  - Встроенные типы
  - Функции
  - Управление потоками
  - Сопрограммы
- R

Давайте решим задачу

- Дан список списков: [[1,2,3,…],[4,5,7,…],…]
- Найти сумму вторых элементов всех вложенных списков: 2+5+…

- Предложите разные решения на Python
- Может быть есть красивые решения на других языках?
- Сколько человек выбрали, то или иное решение?
- Какое решение лучше и почему?

Литература

- Luciano Ramalho. Fluent Python
- Joel Grus. Data Science from Scratch
- Allen B. Downey. Think Complexity
- PEP8
- Matloff, Norman S. The art of R programming: tour of statistical software design

Эскимосы и снег

- Что этот язык может «делать»?

- Программа – последовательность символов, определяющая вычисления
- Язык программирования – это набор правил, определяющих, какие последовательности символов составляют программу и какое вычисление описывает программа

Python

- Интерпретация
- Байт-код
- Встроенные типы реализованы на C
- Динамический
- Мультипарадигменый
- Эталонной реализацией Python является интерпретатор CPython. Распространяется под свободной лицензией Python Software Foundation License. Есть реализация для JVM с возможностью компиляции, CLR. PyPy использует JIT-компиляцию, которая значительно увеличивает скорость выполнения Python-программ.

import this

- Красивое лучше, чем уродливое.
- Явное лучше, чем неявное.
- Простое лучше, чем сложное.
- Сложное лучше, чем запутанное.
- Плоское лучше, чем вложенное.
- Разреженное лучше, чем плотное.
- Читаемость имеет значение.
- Особые случаи не настолько особые, чтобы нарушать правила.
- При этом практичность важнее безупречности.
- Ошибки никогда не должны замалчиваться.
- Если не замалчиваются явно.
- Встретив двусмысленность, отбрось искушение угадать.
- Должен существовать один — и, желательно, *только* один — очевидный способ сделать это.
- Хотя он поначалу может быть и не очевиден, если вы не голландец.
- Сейчас лучше, чем никогда.
- Хотя никогда зачастую лучше, чем *прямо* сейчас.
- Если реализацию сложно объяснить — идея плоха.
- Если реализацию легко объяснить — идея, *возможно*, хороша.
- Пространства имён — отличная вещь! Давайте будем делать их больше!

Метод k ближайших соседей

- Метрика близости
- Голосование k ближайших соседей
- Если результат равный – то убираем самого дальнего соседа или считаем средневзвешенный голос

- Для целей обучения именно программированию не используем **Scikit-learn** или аналоги, хотя можно использовать для сравнения…

Задача

- Давайте познакомимся:
  - Какая у Вас ближайшая станция метро?
  - Что Вы пьете по утрам? Чай или Кофе?
- Разбиваемся на команды 4-5 человек:
  - Распределение ролей
  - Парное программирование
  - Подготовка исходных данных
  - Тестирование
  - Анализ – какое k лучше?
  - Показ решения
- Обсуждение
  - Какое решение лучше и почему?

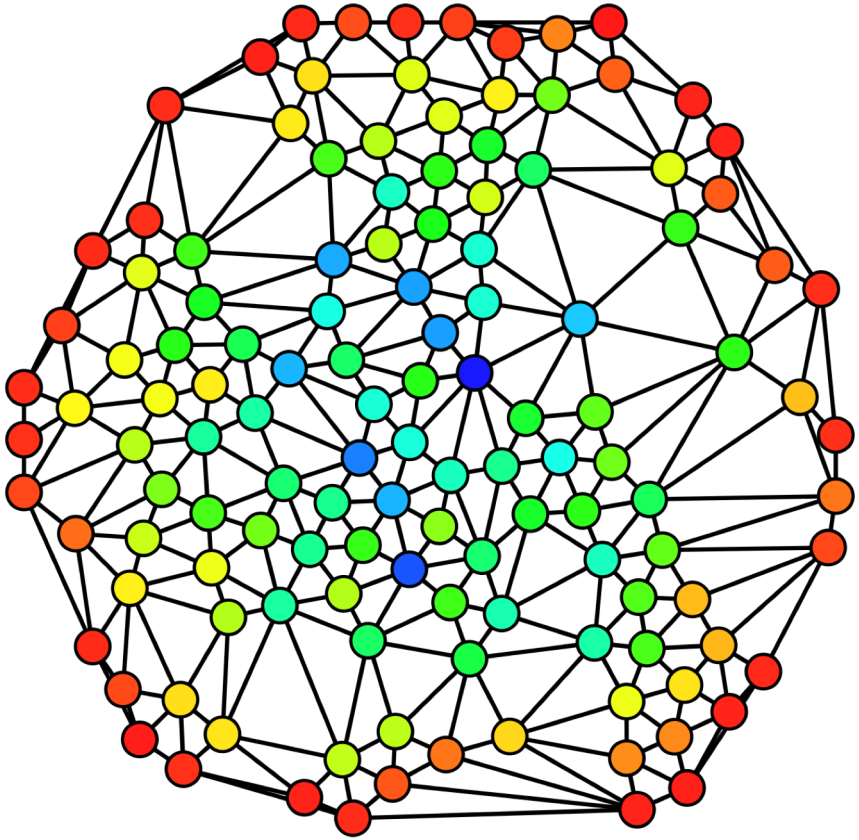
Переменные

- Переменные - это не ящики, это этикетки

Jupiter….

Задача

- Давайте соберем информацию о друзьях из VK
- Оценить «центральность по посредничеству»

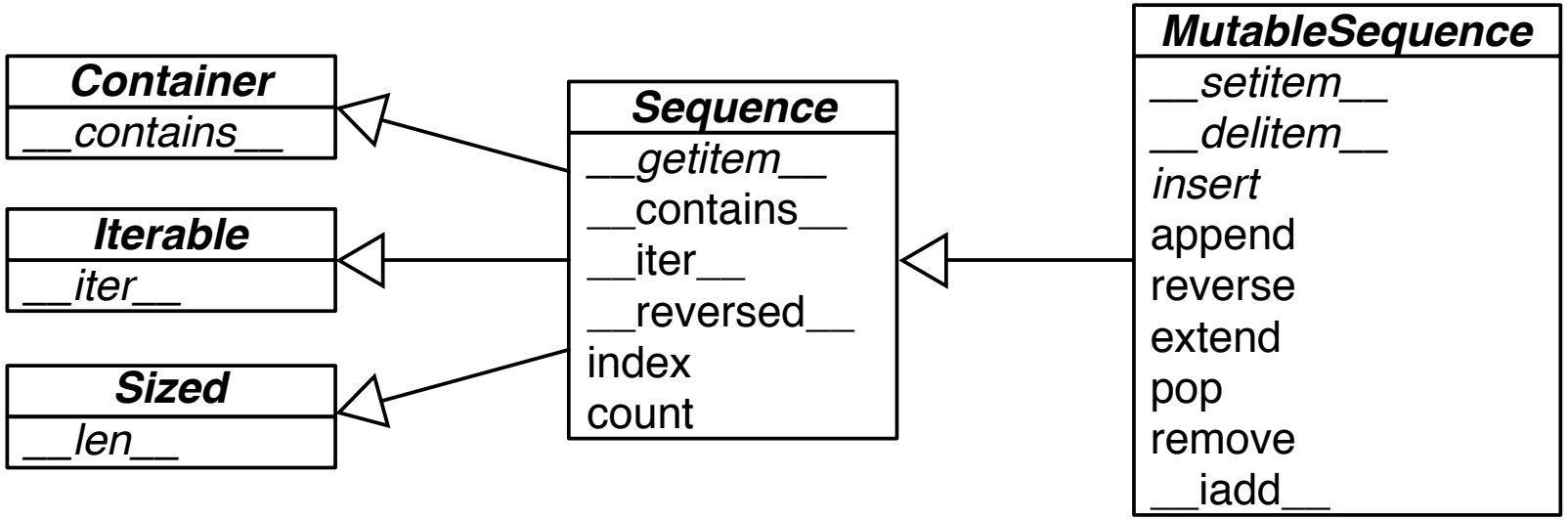


Виды последовательностей

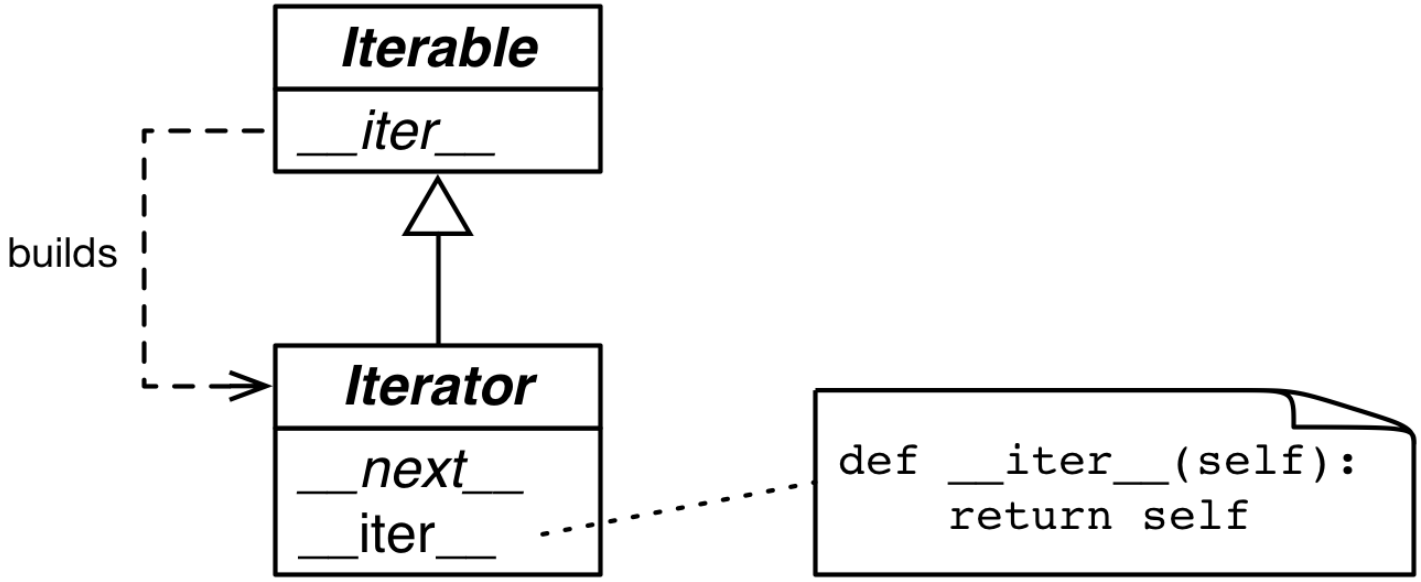
- Контейнерные:
  - list, tuple, collections.deque
- Плоские:
  - str, bytes, bytearray, memoryview, array.array

- Изменяемые:
  - list, bytearray, collections.deque, memoryview, array.array
- Неизменяемые:
  - tuple, str, bytes

Последовательности



Итератор


Iterable
__iter__
builds
Iterator
__next__
__iter__
def __iter__(self):
return self

Итератор

- Решает проблему просмотра данных не помещающихся в память
- Он делает это лениво….
- Итераторы используются для поддержки:
  - for
  - Конструирования коллекций
  - Построчного просмотра файлов
  - Списковых и словарных включений
  - Распаковки кортежей
  - Распаковки фактических параметров *

Контрольная работа

- Реализовать конечный автомат
- Вход: целое число – определяет поведение автомата

При преобразовании в бинарный вид, например числа 1:

0	0	0	0	0	0	0	1
111	110	101	100	011	010	001	000

Верхняя строка показывает какое число нужно записать в среднюю ячейку текущего вектора при данном состоянии смежных трех ячеек (первая и последняя колонка не меняются).

Начальное состояние вектора:

0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

- Выход: последовательность состояний вектора
- Используем чистый Python

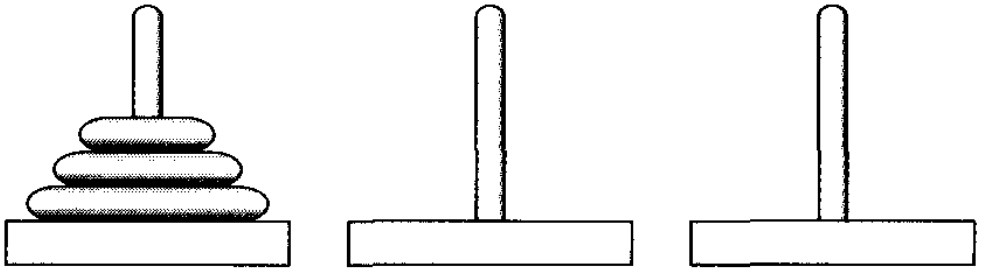
Домашнее задание

- Реализовать модель Шеллинга (модель расовой сегрегации)
- Дан квадрат n x n. 45% клеток синие, 45% клеток красные, 10% клеток пустые. Начальное заполнение в случайном порядке.
- Клетка «счастлива» если у нее 2 или более соседа одного с ней цвета. Соседи – это 8 клеток вокруг данной.
- Моделирование: выбрать случайным образом «несчастную» клетку и переместить ее в случайно выбранную пустую клетку.
- Вывести квадраты через данное некоторое количество шагов иллюстрирующее расовую сегрегацию.

Домашнее задание

1. Дано n отрезков на плоскости. Есть ли среди них пересечения?
2. Дано n точек на плоскости. Построить выпуклый многоугольник, который включает все эти точки.
3. На плоскости дан многоугольник (необязательно выпуклый). Дана точка. Определить принадлежит ли точка многоугольнику.

Контрольная



Даны три стержня, на один из которых нанизаны ***n*** колец, причём кольца отличаются размером и лежат меньшее на большем. Задача состоит в том, чтобы перенести пирамиду на третий стержень. За один раз разрешается переносить только одно кольцо, причём нельзя класть большее кольцо на меньшее.

1. Решить с использованием рекурсии
2. Решить без использования рекурсии

Вход: константа ***n***. Выход - печать вида:

<номер хода>;<номер стержня откуда>;<номер стержня куда>

Второе задание: написать генераторную функцию без аргументов для чисел Фибоначчи: $F_1=1$, $F_2=1$, $F_3=F_1+F_2$ … $F_n=F_{n-2}+F_{n-1}$

Продемонстрировать как вызвать данную генераторную функцию чтобы вернуть первые 5 чисел Фибоначчи

Разбор полетов....

```
def fib():
    old, new = 0, 1
    while True:
        yield old
        old, new = new, old + new
for i in fib():
    print(i)
    if i > 10:
        break
```

0??????????????????????

```
def move(disks, source, temp, target):
    if disks == 1:
        print('Переложили кольцо 1 с стержня {} на стержень {}.'.format(source, target))
        return


    move(disks - 1, source, target, temp)
    print('Переложили кольцо {} с стержня {} на стержень {}.'.format(disks, source, target))
    move(disks - 1, temp, source, target)
```

Разбор полетов....

```
n=int(input())
#n=3
l=1
def moveTower(n, otkuda, kuda, withPole):
    if n >= 1:
        moveTower(n-1, otkuda, withPole, kuda)
        moveDisk(otkuda, kuda)
        global l
        l += 1
        moveTower(n-1, withPole, kuda, otkuda)
def moveDisk(fp,tp):
    #l=1
    print("nomer hoda:", l, ";otkuda:", fp, ";kuda:", tp)
moveTower(n,"A","B","C")
```

ДЗ: Алгоритм ID3

PEP8

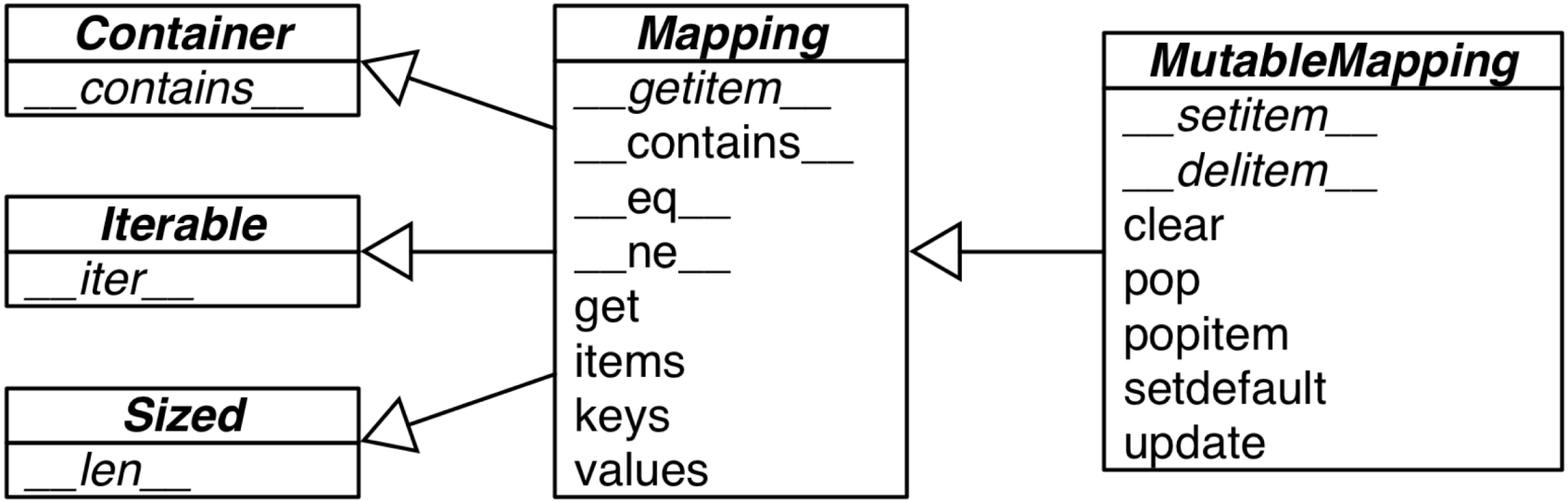
Yes:

```
i = i + 1
submitted += 1
x = x*2 - 1
hypot2 = x*x + y*y
c = (a+b) * (a-b)
```

No:

```
i=i+1
submitted +=1
x = x * 2 - 1
hypot2 = x * x + y * y
c = (a + b) * (a - b)
```

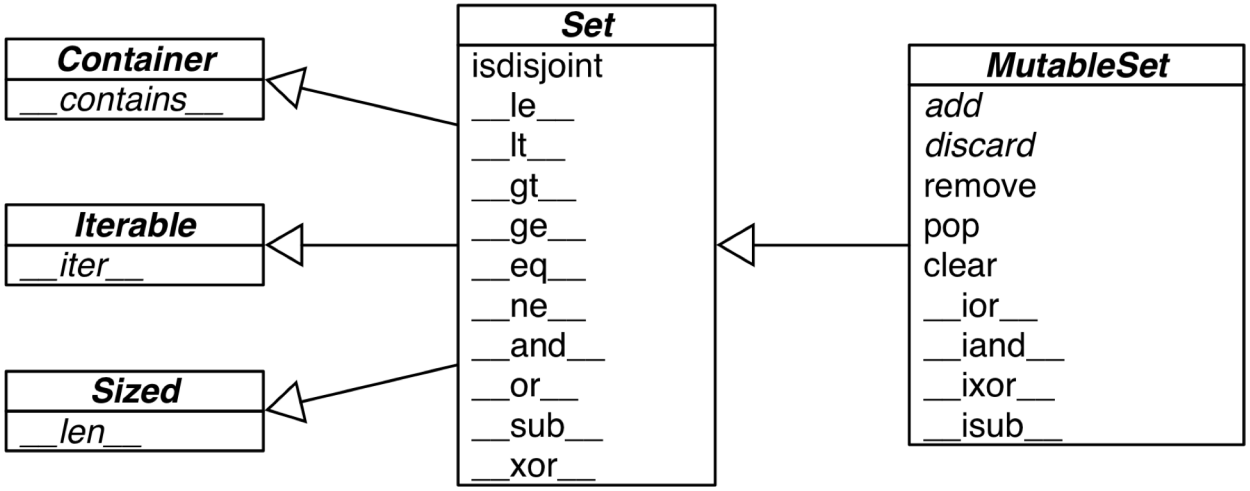
Отображения



Ограничение на ключи

- Ключ должен быть хэшируемым объектом
- Хэшируемый объект – поддерживает:
  - Метод __hash__
  - Метод __eq__
- Если объекты равны, то и их хэш-значения тоже должны быть равны.

Множества


Container
__contains__
Iterable
__iter__
Sized
__len__
Set
isdisjoint
__le__
__lt__
__gt__
__ge__
__eq__
__ne__
__and__
__or__
__sub__
__xor__
MutableSet
add
discard
remove
pop
clear
__ior__
__iand__
__ixor__
__isub__

Операции на множествах

$S \cap Z$	`s & z`	$S \setminus Z$	`s - z`		
	`z & s`		`z - s`	$e \in S$	`e in s`
				$S \subseteq Z$	`s <= z`
	`s &= z`		`s -= z`		
				$S \subset Z$	`s < z`
$S \cup Z$	`s | z`			$S \supseteq Z$	`s >= z`
	`z | s`	$S \Delta Z$	`s ^ z`		
			`z ^ s`	$S \supset Z$	`s > z`
	`s |= z`				
			`s ^= z`		

Контрольная

- Постойте список уникальных типов самолетов зарегистрированных в России
- Какой тип самолета имеет самую раннюю дату выдачи сертификата?
- Постройте запрос: Владелец аэропорта, Аэропорт, Пассажиропоток суммарный за 2018 год, Грузопоток суммарный за 2018 год
- Перечислите аэропорты где пассажиропоток меньше медианы, а грузопоток больше медианы
- Перечислите авиакомпании у которых нет типов воздушных судов зарегистрированных в России
- Выведите список: Месяц, суммарный пассажиропоток за данный месяц, аэропорт в котором пассажиропоток в данном месяце максимальный
- Выведите список: Тип аэропорта, средний грузопоток в месяц в аэропортах данного типа