

A Practical Introduction to Machine Learning in Python

Day 1 - Monday

»Introduction«

Damian Trilling
Anne Kroon

d.c.trilling@uva.nl, @damian0604
a.c.kroon@uva.nl, @annekroon

Gesis

March 9, 2020

Today

① Introducing...

... the people

② Defining CSS

... some definitions

Are we doing Big Data research?

Computational social science

③ CSS project workflow

step-by-step

A good workflow

④ Best practices

Open science

Clean, high-quality code

Exercise

datatypes

Generators

Scaling up

All course materials can be found at...

<https://github.com/annekroon/gesis-ml-learning>

Introducing...

... the people

Introducing... Damian



dr. Damian Trilling
Assistant Professor Political Communication &
Journalism

- interested in political communication and journalism in a changing media environment and in innovative (digital, large-scale, computational) research methods

@damian0604 | d.c.trilling@uva.nl
| www.damiantrilling.net

Introducing... Anne



dr. Anne Kroon

Assistant Professor Corporate Communication

- Research focus on biased AI in recruitment, and media bias regarding minorities
- text analysis using automated approaches, word embeddings

@annekroon |a.c.kroon@uva.nl |<http://www.uva.nl/profiel/k/r/a.c.kroon/a.c.kroon.html>

Short poll

Do you need

- a** an intro
- b** a brief refresher
- c** nothing

on

- i** datatypes (int, float, string, lists, dictionaries)
- ii** control flow statements (for, if, try/except)
- iii** ways to run your code (notebooks vs IDE's vs text editors)

?

We will try to adapt today's programme to your needs!



Dan Ariely

6 januari 2013 · 🌐



Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...



Rachel Daxtor en 2,9 d. anderen

144 opmerkingen 1,3 d. keer gedeeld



Leuk



Opmerking plaatsen



Delen

BIG DATA ANALYTICS

Is Big Data Dying?



By Aravind Sekar — Last updated Nov 30, 2018



Share





What is Big Data?

A simple technical definition could be:

Everything that needs so much computational power and/or storage that you cannot do it on a regular computer.

What is Big Data?

What is Big Data?

Vis, 2013

- “commercial” definition (Gartner): “‘Big data’ is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”

What is Big Data?

Vis, 2013

- boyd & Crawford definition:
 - ① Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
 - ② Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
 - ③ Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

What is Big Data?

Vis, 2013

- “commercial” definition (Gartner): “‘Big data’ is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”
- boyd & Crawford definition:
 - ① Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
 - ② Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
 - ③ Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

Implications & criticism

Implications & criticism

boyd & Crawford, 2012

- ① Big Data changes the definition of knowledge
- ② Claims to objectivity and accuracy are misleading
- ③ Bigger data are not always better data
- ④ Taken out of context, Big Data loses its meaning
- ⑤ Just because it is accessible does not make it ethical
- ⑥ Limited access to Big Data creates new digital divides

APIs, researchers and tools *make* Big Data

APIs, researchers and tools *make* Big Data

Vis, 2013

Inevitable influences of:

- APIs
- filtering, search strings, ...
- changing services over time
- organizations that provide the data

Epistemologies and paradigm shifts

Kitchen, 2014

Epistemologies and paradigm shifts

Kitchin, 2014

- (Reborn) empiricism: purely inductive, correlation is enough

Epistemologies and paradigm shifts

Kitchin, 2014

- (Reborn) empiricism: purely inductive, correlation is enough
- Data-driven science: knowledge discovery guided by theory

Epistemologies and paradigm shifts

Kitchin, 2014

- (Reborn) empiricism: purely inductive, correlation is enough
- Data-driven science: knowledge discovery guided by theory
- Computational social science and digital humanities: employ Big Data research within existing epistemologies
 - DH: descriptive statistics, visualizations
 - CSS: prediction and simulation

Are we doing Big Data research in this course?

Depends on the definition

- Not if we take a definition that *only* focuses on computing power and the amount of data
- **But:** We are using the same techniques. And they *scale* well.

Are we doing Big Data research in this course?

Depends on the definition

- Not if we take a definition that *only* focuses on computing power and the amount of data
- **But:** We are using the same techniques. And they *scale* well.
- Oh, and about that high-performance computing in the cloud: We actually *do* have access to that, so if someone has a really great idea...

Our epistemological underpinnings

Computational Social Science

Computational Social Science

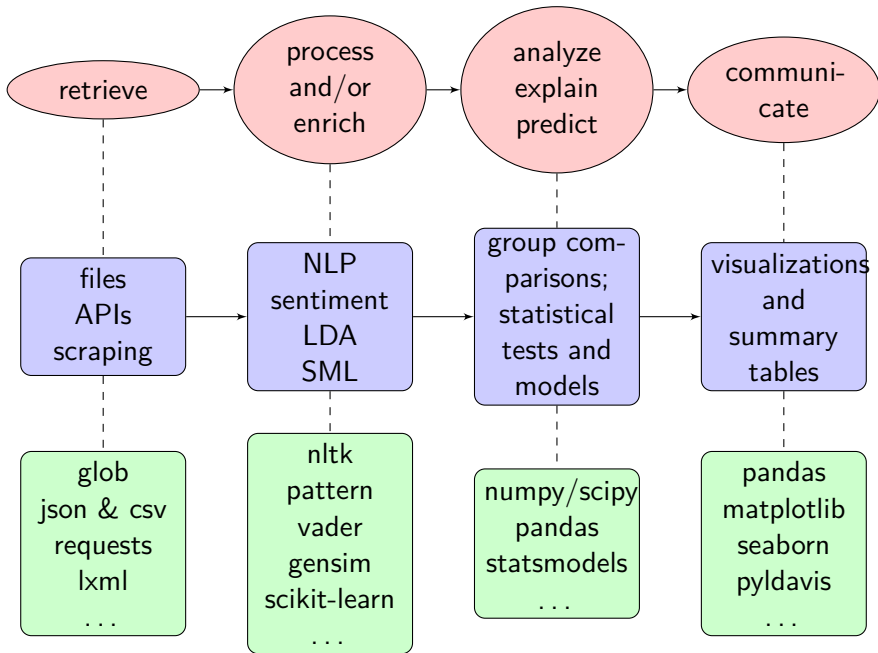
“[...] the computational social sciences employ the scientific method, complementing descriptive statistics with inferential statistics that seek to identify associations and causality. In other words, they are underpinned by an epistemology wherein the aim is to produce sophisticated statistical models that explain, simulate and predict human life.”

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12.
doi:10.1177/2053951714528481

Steps of a CSS project

Different techniques for:

- retrieving data (previous week)
- processing data (previous week)
- analyzing data (main part of this week)
- visualising data (a bit on Friday)



A good workflow

The big picture

Start with pen and paper

① Draw the Big Picture

One script for downloading the data, one script for analyzing

- A set of small navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and other slide controls.

One script for downloading the data, one script for analyzing

- Avoids waste of resources (e.g., unnecessary downloading multiple times)
- Makes it easier to re-use your code or apply it to other data

One script for downloading the data, one script for analyzing

- Avoids waste of resources (e.g., unnecessary downloading multiple times)
- Makes it easier to re-use your code or apply it to other data

Develop components separately

One script for downloading the data, one script for analyzing

- Avoids waste of resources (e.g., unnecessary downloading multiple times)
- Makes it easier to re-use your code or apply it to other data

Start small, then scale up

- Take your plan and solve *one* problem at a time (e.g., parsing a review page; or getting the URLs of all review pages)

Develop components separately

One script for downloading the data, one script for analyzing

- Avoids waste of resources (e.g., unnecessary downloading multiple times)
- Makes it easier to re-use your code or apply it to other data

Start small, then scale up

- Take your plan and solve *one* problem at a time (e.g., parsing a review page; or getting the URLs of all review pages)
- (for instance, by using functions [next slides])

If you copy-paste code, you are doing something wrong

- Write loops!

If you copy-paste code, you are doing something wrong

- Write loops!
- If something takes more than a couple of lines, write a function!

Copy-paste approach (ugly, error-prone, hard to scale up)

```
1 allreviews = []
2
3 response = requests.get('http://xxxxx')
4 tree = fromstring(response.text)
5 reviewelements = tree.xpath('//div[@class="review"]')
6 reviews = [e.text for e in reviewelements]
7 allreviews.extend(reviews)
8
9 response = requests.get('http://yyyyy')
10 tree = fromstring(response.text)
11 reviewelements = tree.xpath('//div[@class="review"]')
12 reviews = [e.text for e in reviewelements]
13 allreviews.extend(reviews)
```

Better: for-loop

(easier to read, less error-prone, easier to scale up (e.g., more URLs, read URLs from a file or existing list))

```
1 allreviews = []
2
3 urls = ['http://xxxxx', 'http://yyyyy']
4
5 for url in urls:
6     response = requests.get(url)
7     tree = fromstring(response.text)
8     reviewelements = tree.xpath('//div[@class="review"]')
9     reviews = [e.text for e in reviewelements]
10    allreviews.extend(reviews)
```

Even better: for-loop with functions
(main loop is easier to read, function can be re-used in multiple contexts)

```
1 def getreviews(url):
2     response = requests.get(url)
3     tree = fromstring(response.text)
4     reviewelements = tree.xpath('//div[@class="review"]')
5     return [e.text for e in reviewelements]
6
7
8 urls = ['http://xxxxx', 'http://yyyyy']
9
10 allreviews = []
11
12 for url in urls:
13     allreviews.extend(getreviews(url))
```

Exercises

Did you pass?

- Think of a way to determine for a list of grades whether they are a pass (>5.5) or fail.
- Can you make that program robust enough to handle invalid input (e.g., a grade as 'ewghjeh')?
- How does your program deal with impossible grades (e.g., 12 or -3)?
- ...

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Advantages

- fast, flexible

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Advantages

- fast, flexible
- allows for nested, unstructured data

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Advantages

- fast, flexible
- allows for nested, unstructured data

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Advantages

- fast, flexible
- allows for nested, unstructured data

Disadvantages

- can be more cumbersome: e.g., inserting a column

Datatypes

Low-level: Native python datatypes

- Booleans, integers, floats, strings, bytes, byte arrays
- Lists, tuples, sets, dictionaries

Advantages

- fast, flexible
- allows for nested, unstructured data

Disadvantages

- can be more cumbersome: e.g., inserting a column
- less consistency checks

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Advantages

- useful convenience functionality, works very intuitively (for tabular data)

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Advantages

- useful convenience functionality, works very intuitively (for tabular data)
- easy, allows for pretty visualization

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Advantages

- useful convenience functionality, works very intuitively (for tabular data)
- easy, allows for pretty visualization

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Advantages

- useful convenience functionality, works very intuitively (for tabular data)
- easy, allows for pretty visualization

Disadvantages

- not suited for one-dimensional or messy / deeply nested data

Datatypes

Higher-level: importing modules

- e.g., numpy, pandas, seaborn

Advantages

- useful convenience functionality, works very intuitively (for tabular data)
- easy, allows for pretty visualization

Disadvantages

- not suited for one-dimensional or messy / deeply nested data
- when your data is very large (machine learning!!)

Datatypes in this course

In this week, we will mainly work with lower-level datatypes (as opposed to, for instance, pandas dataframes)

- Often, ML algorithms require native data types as input (i.e., lists, generators)

Datatypes in this course

In this week, we will mainly work with lower-level datatypes (as opposed to, for instance, pandas dataframes)

- Often, ML algorithms require native data types as input (i.e., lists, generators)
- We have to seriously consider memory:

Datatypes in this course

In this week, we will mainly work with lower-level datatypes (as opposed to, for instance, pandas dataframes)

- Often, ML algorithms require native data types as input (i.e., lists, generators)
- We have to seriously consider memory:
- Maybe size does not apply to your project yet, but in the future you might want to scale up.

Generators

Generators

- We will work with *generators* to deal with memory issues

Generators

Generators

- We will work with *generators* to deal with memory issues
- Generators behave like iterators: loops through elements of an object.

Generators

Generators

- We will work with *generators* to deal with memory issues
- Generators behave like iterators: loops through elements of an object.

Generators

Generators

- We will work with *generators* to deal with memory issues
- Generators behave like iterators: loops through elements of an object.

Behavior of a generator

- Does not hold results in memory

Generators

Generators

- We will work with *generators* to deal with memory issues
- Generators behave like iterators: loops through elements of an object.

Behavior of a generator

- Does not hold results in memory
- Only computes results at the moment you need them (i.e. lazy')

Generators

Generators

- We will work with *generators* to deal with memory issues
- Generators behave like iterators: loops through elements of an object.

Behavior of a generator

- Does not hold results in memory
- Only computes results at the moment you need them (i.e. lazy')
- You can only loop over your object ONCE.

Creating generators: Example 1

```
1 def my_generator(my_list):  
2     for i in my_list:  
3         yield i  
4 example_list = [1, 2, 3, 4]  
5 gen1 = my_generator(example_list)  
6 next(gen1)
```

This will return:

```
1 1
```


Creating generators: Example 2 (shorter)

```
1 my_list = [1,2,3,4]
2 gen = (i for i in my_list)
```

Scaling up

When considering datatypes, consider re-usability, scalability

- Use functions and classes to make code more readable and re-usable
- Avoid re-calculating values
- Think about how to minimize memory usage (e.g., Generators)
- Do not hard-code values, file names, etc., but take them as arguments

Make it robust

You cannot foresee every possible problem.

Most important: Make sure your program does not fail and loose all data just because something goes wrong at case 997/1000.

- Use try/except to explicitly tell the program how to handle errors
- Write data to files (or database) in between
- Use `assert len(x) == len(y)` for sanity checks

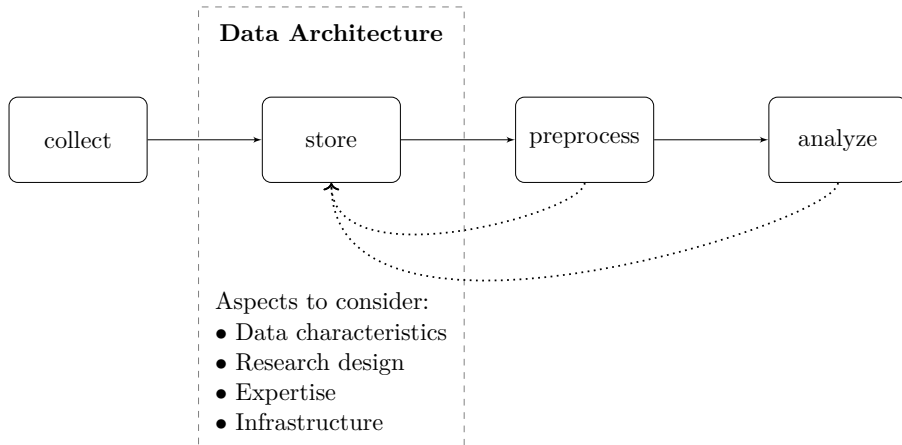
Storing data

Use of databases

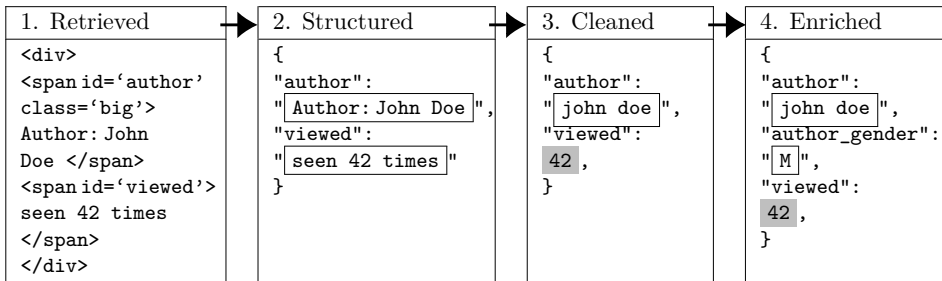
Storing data

- We can store our data in files (often, one CSV or JSON file)
- But that's not very efficient if we have large datasets; especially if we want to select subsets later on
- SQL-databases to store tables (e.g., MySQL)
- NoSQL-databases to store less structured data (e.g., JSON with unknown keys) (e.g., MongoDB, Elasticsearch)
- ⇒ Günther, E., Trilling, D., & Van de Velde, R.N. (2018). But how do we store it? (Big) data architecture in the social-scientific research process. In: *Stuetzer, C.M., Welker, M., & Egger, M. (eds.): Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications*. Cologne, Germany: Herbert von Halem.

Storing data

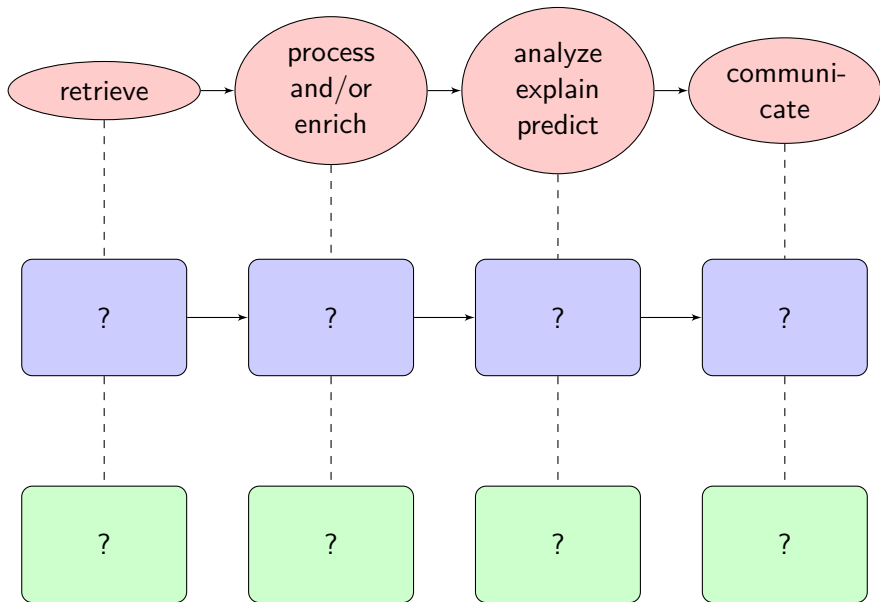


From retrieved data to enriched data



Looking forward

Try to fill in the blanks for your personal CSS project



Long story short:

Don't forget to plan the bigger picture

We will focus on machine learning this week. But for each technique we cover, think about how it fits in *your* workflow.

Long story short:

Don't forget to plan the bigger picture

We will focus on machine learning this week. But for each technique we cover, think about how it fits in *your* workflow. ... and now lets get started!

Types of Automated Content Analysis

	Methodological approach		
	<i>Counting and Dictionary</i>	<i>Supervised Machine Learning</i>	<i>Unsupervised Machine Learning</i>
Typical research interests and content features	visibility analysis sentiment analysis subjectivity analysis	frames topics gender bias	frames topics
Common statistical procedures	string comparisons counting	support vector machines naive Bayes	principal component analysis cluster analysis latent dirichlet allocation semantic network analysis
<div> <div>deductive</div> <div></div> <div>inductive</div> </div>			

Getting started with the IMBD dataset

Backup slides in case we need to do
more fundamentals

Basics]The very, very, basics of programming with Python

The very, very, basics of programming

See also Chapter 4.

Python lingo

Basic datatypes (variables)

int 32

float 1.75

bool True, False

string "Damian"

Python lingo

Basic datatypes (variables)

int 32

float 1.75

bool True, False

string "Damian"

(**variable name** firstname)

"firstname" and **firstname** is not the same.

"5" and **5** is not the same.

But you can transform it: `int("5")` will return 5.

You cannot calculate `3 * "5"` (In fact, you can. It's "555").

But you can calculate `3 * int("5")`

Python lingo

More advanced datatypes

Python lingo

More advanced datatypes

```
list firstnames = ['Damian', 'Lori', 'Bjoern']  
lastnames =  
['Trilling', 'Meester', 'Burscher']
```

Note that the elements of a list, the keys of a dict, and the values of a dict can have any datatype! (Better to be consistent, though!)

Python lingo

More advanced datatypes

```
list firstnames = ['Damian', 'Lori', 'Bjoern']
      lastnames =
      ['Trilling', 'Meester', 'Burscher']
list ages = [18, 22, 45, 23]
```

Note that the elements of a list, the keys of a dict, and the values of a dict can have any datatype! (Better to be consistent, though!)

Python lingo

More advanced datatypes

```
list firstnames = ['Damian', 'Lori', 'Bjoern']
    lastnames =
    ['Trilling', 'Meester', 'Burscher']

list ages = [18, 22, 45, 23]

dict familynames= {'Bjoern': 'Burscher',
                   'Damian': 'Trilling', 'Lori': 'Meester'}

dict {'Bjoern': 26, 'Damian': 31, 'Lori':
     25}
```

Note that the elements of a list, the keys of a dict, and the values of a dict can have any datatype! (Better to be consistent, though!)

Python lingo

Functions

- functions** Take an input and return something else
`int(32.43)` returns the integer 32. `len("Hello")` returns the integer 5.
- methods** are similar to functions, but directly associated with an object. `"SCREAM".lower()` returns the string "scream"

Both functions and methods end with (). Between the (), *arguments* can (sometimes have to) be supplied.

You can write an own function:

```
1 def addone(x):
2     y = x + 1
3     return y
```

Thus, running

```
1 addone(5)
```

returns 6.

Modifying lists and dictionaries

Modifying lists

Appending to a list

```
1 mijnlijst = ["element 1", "element 2"]
2 anotherone = "element 3" # note that this is a string, not a list!
3 mijnlijst.append(anotherone)
4 print(mijnlijst)
```

gives you:

```
1 ["element 1", "element 2", "element 3"]
```

Merging two lists (= extending)

gives you:

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ▶ ↺ 🔍 ↻

Modifying dicts

Adding a key to a dict (or changing the value of an existing key)

```
1 mydict = {"whatever": 42, "something": 11}
2 mydict["somethingelse"] = 76
3 print(mydict)
```

gives you:

```
1 {'whatever': 42, 'somethingelse': 76, 'something': 11}
```

If a key already exists, its value is simply replaced.

Indentation]Indentation: The Python way of structuring your program

Indentation: The Python way of structuring your program

Indentation

Structure

The program is structured by TABs or SPACES

Stack Overflow Developer ...

stackoverflow.com/research/developer-survey-2015

overflow developer survey

stackoverflow

Overview

Developer Profile

Technology

I. Most Popular Technologies

II. Most Loved, Dreaded, and Wanted Tools

III. Desktop Operating System

IV. Text Editor

V. IDE Theme

VI. Source Control

VII. Tabs vs. Spaces

VIII. Caffeine

Work

Community

Back to top

Looking for a job?

about 10% of developers still don't use it.

VII. TABS VS. SPACES

Tabs

45.0%

Spaces

33.6%

It depends

17.0%

Huh?

4.5%

25,807 responses

After millennia of heated debate, mercifully, at long last, we have an answer. **Most developers prefer tabs to spaces.**

Upon closer examination of the data, a trend emerges: Developers increasingly prefer spaces as they gain experience. Stack Overflow reputation correlates with a preference for spaces, too: users who have 10,000 rep or more prefer spaces to tabs at a ratio of 3 to 1.

tab

Alles markeren

Hoofdlettergevoelig

4 van 8 overeenkomsten

Indentation

Structure

The program is structured by TABs or SPACEs

```
1 firstnames=['Damian','Lori','Bjoern']
2 age={'Bjoern': 27, 'Damian': 32, 'Lori': 26}
3 print ("The names and ages of these people:")
4 for naam in firstnames:
5     print (naam,age[naam])
```

Indention

Structure

The program is structured by TABs or SPACEs

```
1 firstnames=['Damian','Lori','Bjoern']
2 age={'Bjoern': 27, 'Damian': 32, 'Lori': 26}
3 print ("The names and ages of these people:")
4 for naam in firstnames:
5     print (naam,age[naam])
```

Don't mix up TABs and spaces! Both are valid, but you have to be consequent!!! Best: always use 4 spaces!

Indentation

Structure

The program is structured by TABs or SPACES

```
1 print ("The names and ages of all these people:")
2 for naam in firstnames:
3     print (naam,age[naam])
4     if naam=="Damian":
5         print ("He teaches this course")
6     elif naam=="Lori":
7         print ("She is a former assistant")
8     elif naam=="Bjoern":
9         print ("He helped teaching this course in the past")
10    else:
11        print ("No idea who this is")
```

Indentation

The line *before* an indented block starts with a *statement* indicating what should be done with the block and ends with a :

Indentation

The line *before* an indented block starts with a *statement* indicating what should be done with the block and ends with a :

Indentation of the block indicates that

Indentation

The line *before* an indented block starts with a *statement* indicating what should be done with the block and ends with a :

Indentation of the block indicates that

- it is to be executed repeatedly (for statement) – e.g., for each element from a list

Indentation

The line *before* an indented block starts with a *statement* indicating what should be done with the block and ends with a :

Indentation of the block indicates that

- it is to be executed repeatedly (for statement) – e.g., for each element from a list
- it is only to be executed under specific conditions (if, elif, and else statements)

Intention

The line *before* an indented block starts with a *statement* indicating what should be done with the block and ends with a :

Indentation of the block indicates that

- it is to be executed repeatedly (for statement) – e.g., for each element from a list
- it is only to be executed under specific conditions (if, elif, and else statements)
- an alternative block should be executed if an error occurs (try and except statements)

Indention

The line *before* an indented block starts with a *statement* indicating what should be done with the block and ends with a :

Indention of the block indicates that

- it is to be executed repeatedly (for statement) – e.g., for each element from a list
- it is only to be executed under specific conditions (if, elif, and else statements)
- an alternative block should be executed if an error occurs (try and except statements)
- a file is opened, but should be closed again after the block has been executed (with statement)

Exercise

We'll now together do some simple exercises ...

Exercises

1. Warming up

- Create a list, loop over the list, and do something with each value (you're free to choose).

2. Did you pass?

- Think of a way to determine for a list of grades whether they are a pass (>5.5) or fail.
- Can you make that program robust enough to handle invalid input (e.g., a grade as 'ewghjeh')?
- How does your program deal with impossible grades (e.g., 12 or -3)?
- ...