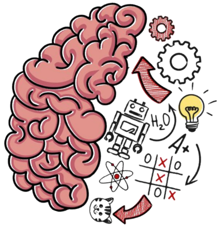
di.unito.it

DIPARTIMENTO DI INFORMATICA

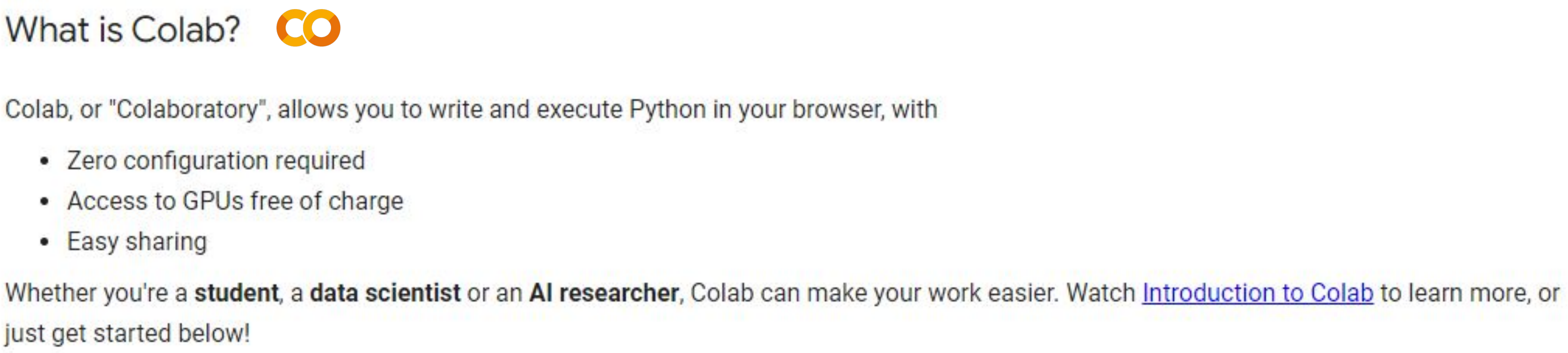
Commission for Orientation and Informatics in schools



#DeactivHate

Artificial Intelligence and
Computational Linguistics
to counter the spread
of hateful messages online

Welcome to Colaboratory!



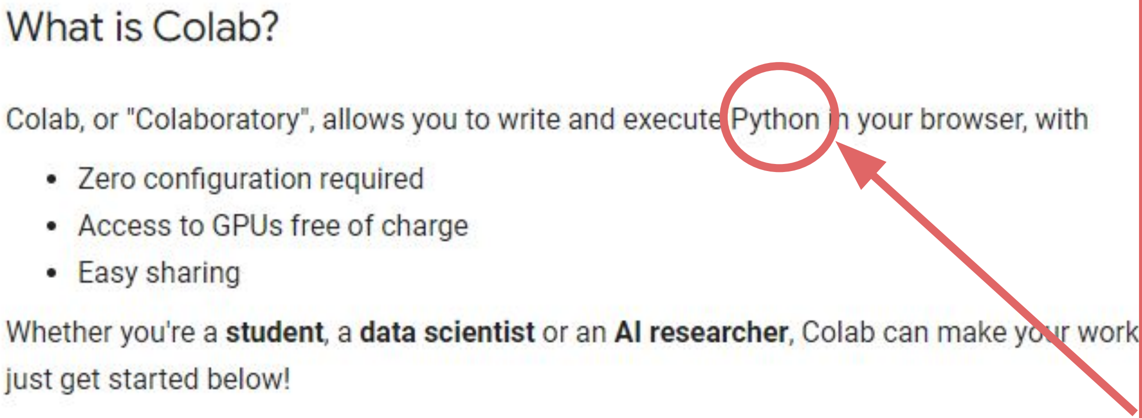
What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

Welcome to Colaboratory!

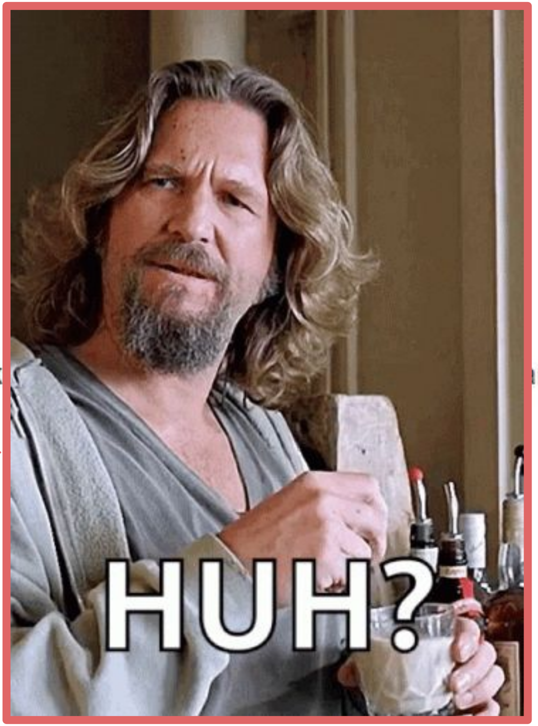


What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work rn more, or just get started below!



Python

Python is a "high-level" object-oriented **programming language**

It is suitable for developing distributed applications, scripting, numerical computation and natural language processing.

Many libraries are available for the most varied uses!

NLTK

scikits learn machine learning in Python

pandas

spaCy

Python Libraries

Something written by **others** that makes **our** work easier!

Introduction

Programming language

- **high-level**: abstracts significantly from the details of a computer's operation and the characteristics of machine language

- **object-oriented**: a programming paradigm through which it is possible to define software objects capable of interacting with each other

Introduction

Programming language

- **is interpreted**: the source code (simple text files with the .py extension) is translated into machine language before being executed by the computer.

- **has various application areas**: web development, database access, desktop applications, games and 3D graphics, scientific and numerical computing, etc...

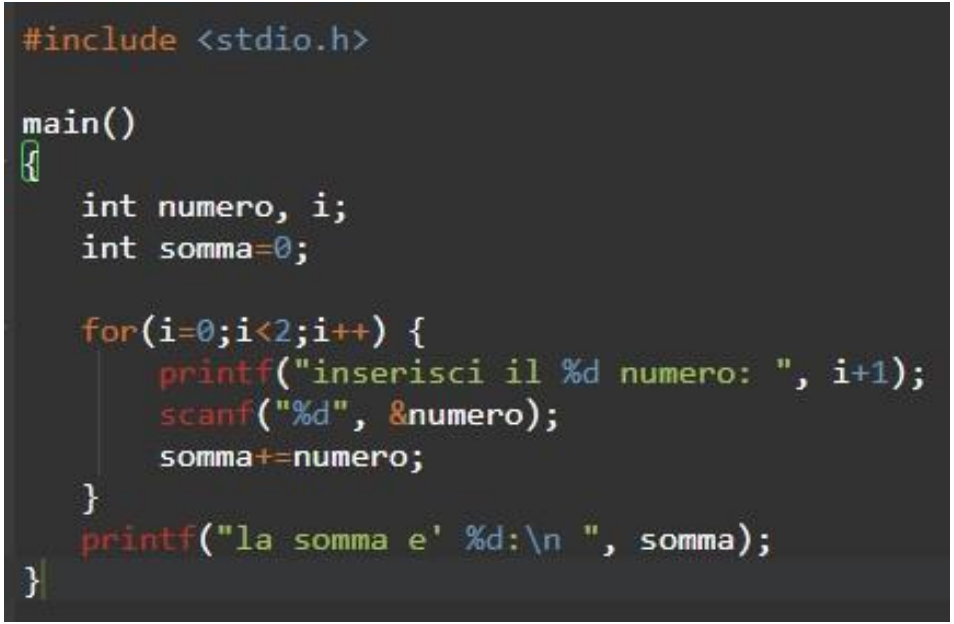
C

1. compiled language
2. not object oriented
3. untyped language:
    a. mandatory variable type declaration
4. no indentation ({code block})
5. ; divides the functions
6. compilation directives
7. array
8. dictionary type absent
9. &&, ||, !

Python

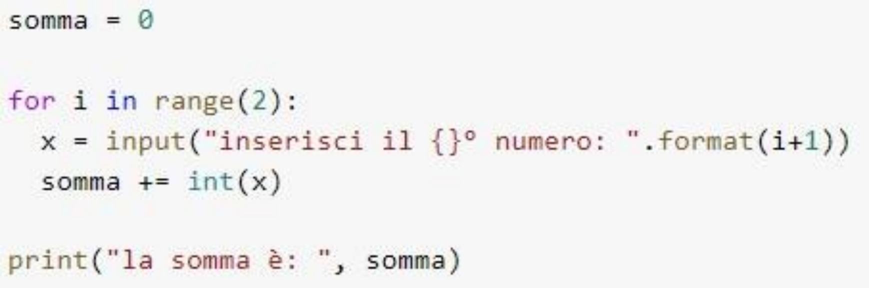
1. interpreted language
2. object oriented
3. typed language:
    a. unnecessary variable type declaration
4. mandatory indentation
5. ; absent
6. libraries
7. list
8. dictionary = {k1:v1,... kn:vn}
9. and, or, not

C



```c
#include <stdio.h>

main()
{
    int numero, i;
    int somma=0;

    for(i=0;i<2;i++) {
        printf("inserisci il %d numero: ", i+1);
        scanf("%d", &numero);
        somma+=numero;
    }
    printf("la somma e' %d:\n ", somma);
}
```

Python

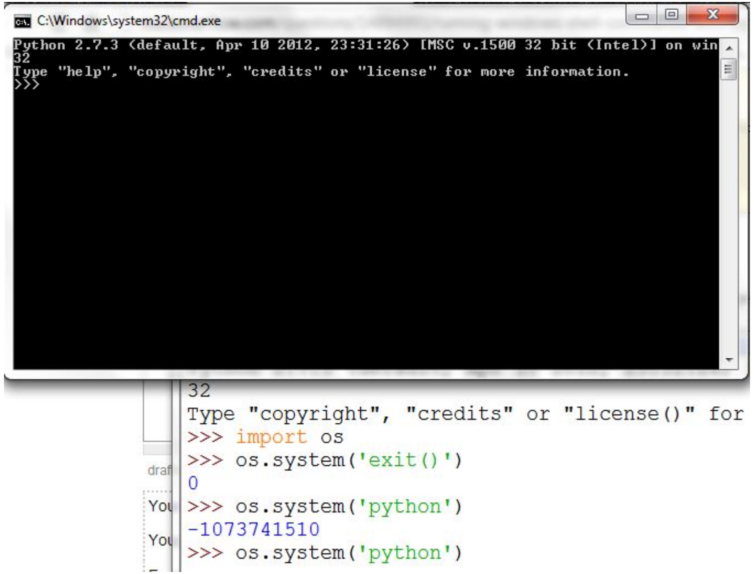


```python
somma = 0

for i in range(2):
  x = input("inserisci il {}° numero: ".format(i+1))
  somma += int(x)

print("la somma è: ", somma)
```

Introduction

It is commonly used to create scripts.

Scripts are nothing more than programs that are designed to run within an operating system shell.



```
Python 2.7.3 (default, Apr 10 2012, 23:31:26) [MSC v.1500 32 bit (Intel)] on win
32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

```
32
Type "copyright", "credits" or "license()" for
>>> import os
>>> os.system('exit()')
0
>>> os.system('python')
-1073741510
>>> os.system('python')
```

Introduction

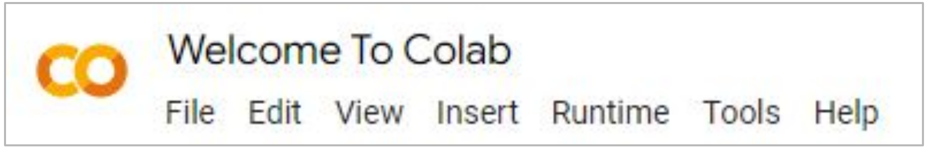
We will use it for:

- analyze (linguistic) data,
- process them
- and classify them

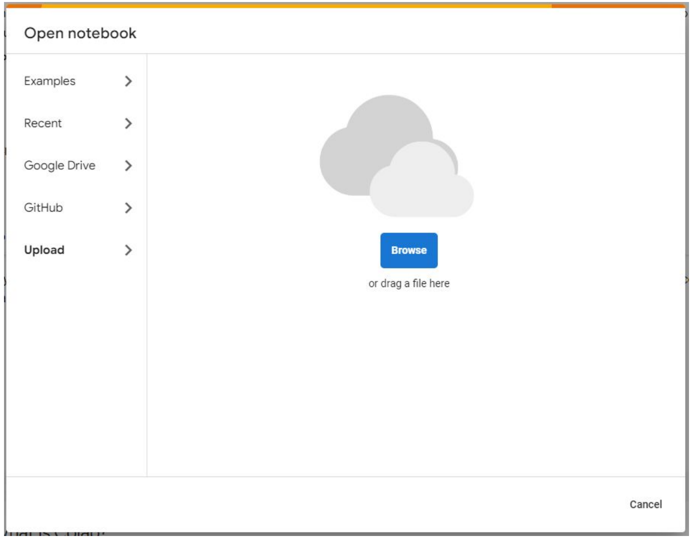
...on Colab

Let’s code !!

1. Access Colab via the link https://colab.research.google.com/
2. Log in via a Gmail based address (i.e. your school email address is fine)
3. Import the available project into your teaching material:



File › Upload Notebook › Browse



Text treatment

1. Cleaning the text
2. Segmentation of a text into sentences
3. Tokenization
4. Normalizing the text
5. Distribution and Relevance of words
6. Extraction of bigrams and trigrams

Cleaning the text

1. remove **@user** from the text

   @user @user Is the repatriation of criminals waging war on the poor?
   → Is repatriating criminals waging war on the poor?

2. replace the url with the **URL** label

   until the intervention of the carabinieri http://www.corrierealtomilanese.com/
   → until the intervention of the police URL

3. convert all words to **lowercase**

   In Magenta clashes between police forces and migrants
   → Magenta clashes with migrant law enforcement forces

Segmentation of a text into sentences

"sentence1 sentence2 … sentenceN" --> ["sentence1", "sentence2", … "sentenceN"]

“Peter Piper picked a peck of pickled peppers. A peck of pickled peppers Peter Piper picked. If Peter Piper picked a peck of pickled peppers, where’s the peck of pickled peppers Peter Piper picked?”

sentence 1: Peter Piper picked a peck of pickled peppers.

sentence 2: A peck of pickled peppers Peter Piper picked.

sentence 3: If Peter Piper picked a peck of pickled peppers,

sentence 4: where’s the peck of pickled peppers Peter Piper picked?

Tokenization

- splitting the text into tokens

  sentence: “the world is round!”
  tokenized sentence: [“the”, “world”, “is”, “round”]

- vocabulary building:
  - eliminate stopwords
  - remove punctuation

Text normalization

- Stemming: Reducing the inflected form of a word to its root form

Example in Spanish:

word form --> stem

"gato" --> "gat-"
"gatos" --> "gat-"

"gata" --> "gat-"
"gatas" --> "gat-"

Text normalization

- Lemmatization: normalization of the inflected form of a word with its respective lemma

Example in Spanish:

word form --> lemma

"gato" --> "gato"

"gatos" --> "gato"

"gata" --> "gato"

"gatas" --> "gato"

Distribution and Relevance of words

- Distribution: number of occurrences of a word

“Peter Piper picked a peck of pickled peppers. A peck of pickled peppers Peter Piper picked. If Peter Piper picked a peck of pickled peppers, where’s the peck of pickled peppers Peter Piper picked?”

n# word “peppers” : 4

Distribution and Relevance of words

- **TF-IDF** (Term Frequency * Inverse Document Frequency): measures the importance of a term with respect to a document or a collection of documents

- TF = number of occurrences of a word w / total number of words in the text
- IDF = log(total number of sentences/number of sentences with the word w)

“Peter Piper picked a peck of pickled peppers. A peck of pickled peppers Peter Piper picked. If Peter Piper picked a peck of pickled peppers, where’s the peck of pickled peppers Peter Piper picked? The end.”

TF of the word "peppers" = 4/35 = 0.11

IDF of the word "peppers" = log(4/5) = 0.80

TF-IDF = 0.11*0.80 = 0.088

Extraction of n-grams

Example: NASA records the sound of wind on Mars for the first time

Bag of Words

1. **Unigrams: the text is split into single word tokens**

   [NASA, records, the, sound, of, wind, on, Mars, for, the, first, time]

2. **N-grams (from 1 to 2), the text is divided into single-word tokens and 2-word pairs**

   [NASA, records, the, sound, of, wind, on, Mars, for, the, first, time, NASA records, records the, the sound, sound of, of wind, wind on, on Mars, Mars for, for the, the first, first time]