

# Lab 2

## String operations ,text extraction

```
#!/homebooks/Untitled2.ipynb

Jupyter Untitled2 Last Checkpoint: 1 hour ago
File Edit View Run Kernel Settings Help
+ X K □ ▶ ■ □ ⇄ Code ▾

JupyterLab Python 3 (ipykernel)

[1]: s= "hello, World"
    print(len(s))
    12

[2]: a="hello"
    b="world"
    result= a + " " + b
    print(result)
    hello world

[3]: !pip install jieba
    Collecting jieba
    Downloading jieba-0.42.1.tar.gz (19.2 MB)
    ----- 19.2/19.2 MB 11.3 MB/s eta 0:00:00M eta 0:00:010:01:01
    Preparing metadata (setup.py) ... done
    Building wheels for collected packages: jieba
    Building wheel for jieba (setup.py) ... done
    Created wheel for jieba: filename=jieba-0.42.1-py3-none-any.whl size=19314459 sha256=39894e4e8f97a65678836db6e15c8a4414b98df85249769ccdaf3bdf3bed248
    Stored in directory: /home/matlab/.cache/pip/wheels/88/a1/a3/5c8ac57cc2f5782ffffc34c95c57c8e5ecb3063dc69541ee7c
    Successfully built jieba
    Installing collected packages: jieba
    Successfully installed jieba-0.42.1

[4]: import jieba

[5]: from urllib import request

[6]: url = "https://www.gutenberg.org/cache/epub/76583/pg76583.txt"

[7]: response = request.urlopen(url)

[8]: con = response.read().decode('utf8')

[22]: import nltk
    nltk.download('punkt')
    [nltk_data] Downloading package punkt to /home/matlab/nltk data...
    [nltk_data] Unzipping tokenizers/punkt.zip.

[22]: True
```

```
Jupyter Untitled2 Last Checkpoint: 1 hour ago
File Edit View Run Kernel Settings Help
+ X K □ ▶ ■ □ ⇄ Code ▾

JupyterLab Python 3 (ipykernel)

[7]: response = request.urlopen(url)

[8]: con = response.read().decode('utf8')

[22]: import nltk
    nltk.download('punkt')
    [nltk_data] Downloading package punkt to /home/matlab/nltk data...
    [nltk_data] Unzipping tokenizers/punkt.zip.

[22]: True

[23]: from nltk.tokenize import word_tokenize
    url = "https://www.gutenberg.org/cache/epub/76583/pg76583.txt"
    response = request.urlopen(url)
    tokens = word_tokenize(con)

[23]: !pip install nltk
    Requirement already satisfied: nltk in /home/software/software/lib/python3.12/site-packages (3.8.1)
    Requirement already satisfied: click in /home/software/software/lib/python3.12/site-packages (from nltk) (8.1.7)
    Requirement already satisfied: joblib in /home/software/software/lib/python3.12/site-packages (from nltk) (1.4.2)
    Requirement already satisfied: regex-2021.8.3 in /home/software/software/lib/python3.12/site-packages (from nltk) (2023.10.3)
    Requirement already satisfied: tqdm in /home/software/software/lib/python3.12/site-packages (from nltk) (4.66.4)

[24]: print(tokens[:50])

['\u00ffThe', 'Project', 'Gutenberg', 'ebook', 'of', 'the', 'man', 'who', 'mastered', 'time', 'This', 'ebook', 'is', 'for', 'the', 'use',
'of', 'anyone', 'anywhere', 'in', 'the', 'United', 'States', 'and', 'most', 'other', 'parts', 'of', 'the', 'world', 'at', 'no', 'cost', 'an
d', 'with', 'almost', 'no', 'restrictions', 'whatsoever', 'You', 'may', 'copy', 'it', 'give', 'it', 'away', 'or', 're-use', 'it',
'under', 'the', 'terms', 'of', 'the', 'Project', 'Gutenberg', 'License', 'included', 'with', 'this', 'ebook', 'or', 'online', 'at', 'www.gu
tenberg.org', 'If', 'you', 'are', 'not', 'located', 'in', 'the', 'United', 'States', 'you', 'will', 'have', 'to', 'check', 'the',
'laws', 'of', 'the', 'country', 'where', 'you', 'are', 'located', 'before', 'using', 'this', 'ebook', 'Title', 'The', 'man', 'wh
o', 'mastered', 'time', 'Author', 'Ray', 'Cummings', 'Release', 'date', 'July', '14', '1925', 'ebook', '76583',
'Language', 'English', 'Original', 'publication', 'New', 'York', 'NY', 'Ace', 'Books', '1929', 'Credits',
'Greg', 'Weeks', 'Paul', 'Ereaut', 'Mary', 'Meenan', 'the', 'Online', 'Distributed', 'Proofreading']
```

## Documentation:-

### Source of the Website

The data is taken from Project Gutenberg (<https://www.gutenberg.org>), a well-known website that provides free access to public domain books. Specifically, I used this link: <https://www.gutenberg.org/cache/epub/76503/pg76503.txt>, which hosts the plain text version of the book *"The Man Who Mastered Time."*

### What I'm Doing with the Book and Why

I'm taking the text from the book so I can practice programming. Basically, I want to:

- Learn how language is structured and how to break it down.
- Split the book into words/tokens
- See how computers read and work with text,

I'm not changing, copying, or selling the book—it's just for my own learning and understanding.

### Copyrights and licenses

#### Most permissions not needed

Most permission requests we receive do not require a custom response. The vast majority of Project Gutenberg eBooks are in the public domain in the US. This means that nobody can grant, or withhold, permission to do with this item as you please.

"As you please" includes any commercial use, republishing in any format, making derivative works or performances, etc. Read more about the public domain in [Wikipedia](#).

#### Linking

No permission is needed to link to [www.gutenberg.org](http://www.gutenberg.org) or to any page or address within it.

To link to a specific eBook, link to the landing page for that eBook, not to specific files. So, for example, you would link to [www.gutenberg.org/ebooks/19033](http://www.gutenberg.org/ebooks/19033) rather than "deep linking" to the specific HTML, plain text, etc. [www.gutenberg.org](http://www.gutenberg.org) prohibits deep links automatically (based on the HTTP referrer), because from time to time the underlying structure of an eBook might change.

