Introduction to Natural Language Processing using spaCy

Dr. Bambang Purnomosidi D. P.

Faculty of Information Technology

Universitas Teknologi Digital Indonesia

bpdp@utdi.ac.id

Note

1. In ths guide, “**$**” is a prompt for shell (Linux). If you use Windows, you may need toactivate PowerShell or Command prompt (C:\Path\Whatever>). This prompt should not be typed.
2. Any command which should be typed inside the box, will be printed bold, size 14, blue colour. For example:

```
$ micromamba create -n py312-nlp python=3.12
conda-forge/noarch                                  18.4MB @   1.5MB/s
13.7s
conda-forge/linux-64                                41.1MB @   2.7MB/s
18.8s

error  libmamba Could not lock non-existing path
'/home/bpdp/.mamba/pkgs'

Transaction

  Prefix:
/home/bpdp/software/python-dev-tools/micromamba-root/envs/py312-nlp

  Updating specs:

   - python=3.12
...
...
```

1. Preparation

There are some software which need to be installed first:

1. Python, installed by Micromamba.
2. spaCy and spaCy trained model - pipeline for specific language.
3. JupyterLab

1.1 Install Micromamba

See the URL: https://mamba.readthedocs.io/en/latest/installation/micromamba-installation.html

If you want to take a look at Micromamba releases, head over to https://github.com/mamba-org/micromamba-releases/releases - you can have an installer for your OS on that page. To check whether you have Micromamba properly installed on your OS:

```
$ micromamba
Version: 2.0.2

Usage:
/home/bpdp/software/python-dev-tools/micromamba-2.0.2-0/bin/micromamba
[OPTIONS] [SUBCOMMAND]

Options:
  -h,--help                    Print this help message and exit
  --version


Configuration options:
  --rc-file TEXT ...           Paths to the configuration files to use
  --no-rc                      Disable the use of configuration files
  --no-env                     Disable the use of environment variables


Global options:
  -v,--verbose                 Set   verbosity   (higher   verbosity   with
multiple -v, e.g. -vvv)
                          --log-level            ENUM:value             in
{critical->5,debug->1,error->4,info->2,off->6,trace->0,warning->3}       OR
{5,1,4,2,6,0,3}
                               Set the log level
  -q,--quiet                   Set quiet mode (print less output)
  -y,--yes                     Automatically   answer   yes   on   prompted
questions
  --json                       Report all output as json
  --offline                    Force use cached repodata
  --dry-run                    Only display what would have been done
  --download-only              Only  download  and  extract  packages,  do  not
link them into environment.
  --experimental               Enable experimental features


Prefix options:
  -r,--root-prefix TEXT        Path to the root prefix
  -p,--prefix TEXT             Path to the target prefix
  --relocate-prefix TEXT       Path to the relocation prefix
  -n,--name TEXT               Name of the target prefix

Subcommands:
  shell                        Generate shell init scripts
  create                       Create new environment
  install                      Install packages in active environment
  update                         Update packages in active environment
  self-update                    Update micromamba
  repoquery                      Find   and   analyze   packages   in   active
environment or channels
  remove                         Remove packages from active environment
  list                           List packages in active environment
  package                        Extract a package or bundle files into an
archive
  clean                          Clean package cache
  config                         Configuration of micromamba
  info                           Information about micromamba
  constructor                    Commands  to  support  using  micromamba  in
constructor
  env                            List environments
  activate                       Activate an environment
  run                            Run an executable in an environment
  ps                             Show, inspect or kill running processes
  auth                           Login or logout of a given host
  search                         Find  packages  in  active  environment  or
channels
                                 This  is  equivalent  to  `repoquery  search`
command
$
```

1.2 Create an Environment

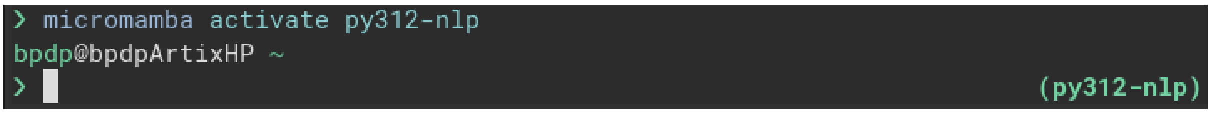
We can have more than 1 Python version installation using *micromamba*. Every installation is called an **environment**. All of those environments do not interfere with each other. For this workshop, create an environment:

```
$ micromamba create -n py312-nlp python=3.12
conda-forge/noarch                                   18.4MB @   1.5MB/s 13.7s
conda-forge/linux-64                                 41.1MB @   2.7MB/s 18.8s

error libmamba Could not lock non-existing path '/home/bpdp/.mamba/pkgs'

Transaction

  Prefix:
/home/bpdp/software/python-dev-tools/micromamba-root/envs/py312-nlp

  Updating specs:

   - python=3.12


  Package                  Version  Build                  Channel
Size
────────────────────────────────────────────────────────────────────────────
───
  Install:
────────────────────────────────────────────────────────────────────────────
───
```

```
  + _libgcc_mutex               0.1  conda_forge          conda-forge  Cached
  + _openmp_mutex               4.5  2_gnu                conda-forge  Cached
  + bzip2                 1.0.8  h4bc722e_7       conda-forge  Cached
...
...
...
  Summary:

  Install: 25 packages

  Total download: 4MB

───────────────────────────────────────────────────────────────────────────
───


Confirm changes: [Y/n] Y

Transaction starting
pip                                                  1.2MB @   1.4MB/s  0.8s
openssl                                              2.9MB @   2.7MB/s  1.0s
Linking _libgcc_mutex-0.1-conda_forge
Linking ld_impl_linux-64-2.43-h712a8e2_2
Linking ca-certificates-2024.12.14-hbcca054_0
Linking libgomp-14.2.0-h77fa898_1
Linking _openmp_mutex-4.5-2_gnu
Linking libgcc-14.2.0-h77fa898_1
Linking openssl-3.4.0-h7b32b05_1
Linking libzlib-1.3.1-hb9d3cd8_2
Linking liblzma-5.6.3-hb9d3cd8_1
Linking libgcc-ng-14.2.0-h69a702a_1
Linking libexpat-2.6.4-h5888daf_0
Linking libsqlite-3.47.2-hee588c1_0
Linking libffi-3.4.2-h7f98852_5
Linking tk-8.6.13-noxft_h4845f30_101
Linking libxcrypt-4.4.36-hd590300_1
Linking bzip2-1.0.8-h4bc722e_7
Linking ncurses-6.5-he02047a_1
Linking libuuid-2.38.1-h0b41bf4_0
Linking libnsl-2.0.1-hd590300_0
Linking readline-8.2-h8228510_1
Linking tzdata-2024b-hc8b5060_0
Linking python-3.12.8-h9e4cc4f_1_cpython
Linking wheel-0.45.1-pyhd8ed1ab_1
Linking setuptools-75.6.0-pyhff2d567_1
Linking pip-24.3.1-pyh8b19718_2

Transaction finished

To activate this environment, use:

      micromamba activate py312-nlp

Or to execute a single command in this environment, use:

      micromamba run -n py312-nlp mycommand

$
```

1.3 Activate The Environment

```
$ micromamba activate py312-nlp
$
```



```
❯ micromamba activate py312-nlp
bpdp@bpdpArtixHP ~
❯                                                          (py312-nlp)
```

1.4 Install spaCy

```
$ pip install spacy
Collecting spacy
                                                                Downloading
spacy-3.8.3-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.meta
data (27 kB)
Collecting spacy-legacy<3.1.0,>=3.0.11 (from spacy)
  Downloading spacy_legacy-3.0.12-py2.py3-none-any.whl.metadata (2.8 kB)
Collecting spacy-loggers<2.0.0,>=1.0.0 (from spacy)
  Downloading spacy_loggers-1.0.5-py3-none-any.whl.metadata (23 kB)
...
...
Downloading markdown_it_py-3.0.0-py3-none-any.whl (87 kB)
Downloading pygments-2.19.1-py3-none-any.whl (1.2 MB)
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
1.2/1.2 MB 3.2 MB/s eta 0:00:00
Downloading
wrapt-1.17.0-cp312-cp312-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2
_17_x86_64.manylinux2014_x86_64.whl (89 kB)
Downloading mdurl-0.1.2-py3-none-any.whl (10.0 kB)
Installing   collected   packages:   cymem,   wrapt,   wasabi,   urllib3,
typing-extensions,   tqdm,   spacy-loggers,   spacy-legacy,   shellingham,
pygments,  packaging,  numpy,  murmurhash,  mdurl,  MarkupSafe,  marisa-trie,
idna,   cloudpathlib,   click,   charset-normalizer,   certifi,   catalogue,
annotated-types,  srsly,  smart-open,  requests,  pydantic-core,  preshed,
markdown-it-py,  language-data,  jinja2,  blis,  rich,  pydantic,  langcodes,
typer, confection, weasel, thinc, spacy
```

```
Successfully installed MarkupSafe-3.0.2 annotated-types-0.7.0 blis-1.1.0
catalogue-2.0.10 certifi-2024.12.14 charset-normalizer-3.4.1 click-8.1.8
cloudpathlib-0.20.0 confection-0.1.5 cymem-2.0.10 idna-3.10 jinja2-3.1.5
langcodes-3.5.0 language-data-1.3.0 marisa-trie-1.2.1 markdown-it-py-3.0.0
mdurl-0.1.2 murmurhash-1.0.11 numpy-2.2.1 packaging-24.2 preshed-3.0.9
pydantic-2.10.4 pydantic-core-2.27.2 pygments-2.19.1 requests-2.32.3
rich-13.9.4 shellingham-1.5.4 smart-open-7.1.0 spacy-3.8.3
spacy-legacy-3.0.12 spacy-loggers-1.0.5 srsly-2.5.0 thinc-8.3.3 tqdm-4.67.1
typer-0.15.1 typing-extensions-4.12.2 urllib3-2.3.0 wasabi-1.1.3
weasel-0.4.1 wrapt-1.17.0
$
```

Check spaCy installation result:

```
$ pip list
Package            Version
------------------ ----------
annotated-types    0.7.0
blis               1.1.0
catalogue          2.0.10
certifi            2024.12.14
charset-normalizer 3.4.1
click              8.1.8
cloudpathlib       0.20.0
confection         0.1.5
cymem              2.0.10
idna               3.10
Jinja2             3.1.5
langcodes          3.5.0
language_data      1.3.0
marisa-trie        1.2.1
markdown-it-py     3.0.0
MarkupSafe         3.0.2
mdurl              0.1.2
murmurhash         1.0.11
numpy              2.2.1
packaging          24.2
pip                24.3.1
preshed            3.0.9
pydantic           2.10.4
pydantic_core      2.27.2
Pygments           2.19.1
requests           2.32.3
rich               13.9.4
setuptools         75.6.0
shellingham        1.5.4
smart-open         7.1.0
spacy              3.8.3
spacy-legacy       3.0.12
spacy-loggers      1.0.5
srsly              2.5.0
```

```
thinc              8.3.3
tqdm               4.67.1
typer              0.15.1
typing_extensions  4.12.2
urllib3            2.3.0
wasabi             1.1.3
weasel             0.4.1
wheel              0.45.1
wrapt              1.17.0
$
```

1.5 Install spaCy Model and Pipeline for English Language

Every language needs their own trained model and pipeline. English is an officially supported language but still lacking for some language - Bahasa Indonesia and Bahasa Malaysia are two examples of unavailable trained models and pipelines.

```
$ spacy download en_core_web_lg
Collecting en-core-web-lg==3.8.0
  Downloading
https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-
3.8.0/en_core_web_lg-3.8.0-py3-none-any.whl (400.7 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
400.7/400.7 MB 3.6 MB/s eta 0:00:00
Installing collected packages: en-core-web-lg
Successfully installed en-core-web-lg-3.8.0
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
$
```

1.6 Install JupyterLab

```
$ pip install jupyterlab
Collecting jupyterlab
  Downloading jupyterlab-4.3.4-py3-none-any.whl.metadata (16 kB)
Collecting async-lru>=1.0.0 (from jupyterlab)
  Downloading async_lru-2.0.4-py3-none-any.whl.metadata (4.5 kB)
Collecting httpx>=0.25.0 (from jupyterlab)
  Downloading httpx-0.28.1-py3-none-any.whl.metadata (7.1 kB)
...
...
...
Downloading pycparser-2.22-py3-none-any.whl (117 kB)
Downloading types_python_dateutil-2.9.0.20241206-py3-none-any.whl (14 kB)
Installing collected packages: webencodings, wcwidth, pure-eval,
ptyprocess, fastjsonschema, websocket-client, webcolors, uri-template,
types-python-dateutil, traitlets, tornado, tinycss2, soupsieve, sniffio,
six, send2trash, rpds-py, rfc3986-validator, pyzmq, pyyaml,
python-json-logger, pycparser, psutil, prompt_toolkit, prometheus-client,
platformdirs, pexpect, parso, pandocfilters, overrides, nest-asyncio,
mistune, jupyterlab-pygments, jsonpointer, json5, h11, fqdn, executing,
defusedxml, decorator, debugpy, bleach, babel, attrs, async-lru, asttokens,
terminado, stack_data, rfc3339-validator, referencing, python-dateutil,
matplotlib-inline, jupyter-core, jedi, httpcore, comm, cffi,
beautifulsoup4, anyio, jupyter-server-terminals, jupyter-client,
jsonschema-specifications, ipython, httpx, arrow, argon2-cffi-bindings,
jsonschema, isoduration, ipykernel, argon2-cffi, nbformat, nbclient,
jupyter-events, nbconvert, jupyter-server, notebook-shim,
jupyterlab-server, jupyter-lsp, jupyterlab
Successfully installed anyio-4.8.0 argon2-cffi-23.1.0
argon2-cffi-bindings-21.2.0 arrow-1.3.0 asttokens-3.0.0 async-lru-2.0.4
attrs-24.3.0 babel-2.16.0 beautifulsoup4-4.12.3 bleach-6.2.0 cffi-1.17.1
comm-0.2.2 debugpy-1.8.11 decorator-5.1.1 defusedxml-0.7.1 executing-2.1.0
fastjsonschema-2.21.1 fqdn-1.5.1 h11-0.14.0 httpcore-1.0.7 httpx-0.28.1
ipykernel-6.29.5 ipython-8.31.0 isoduration-20.11.0 jedi-0.19.2
json5-0.10.0 jsonpointer-3.0.0 jsonschema-4.23.0
jsonschema-specifications-2024.10.1 jupyter-client-8.6.3 jupyter-core-5.7.2
jupyter-events-0.11.0 jupyter-lsp-2.2.5 jupyter-server-2.15.0
jupyter-server-terminals-0.5.3 jupyterlab-4.3.4 jupyterlab-pygments-0.3.0
jupyterlab-server-2.27.3 matplotlib-inline-0.1.7 mistune-3.1.0
nbclient-0.10.2 nbconvert-7.16.5 nbformat-5.10.4 nest-asyncio-1.6.0
notebook-shim-0.2.4 overrides-7.7.0 pandocfilters-1.5.1 parso-0.8.4
pexpect-4.9.0 platformdirs-4.3.6 prometheus-client-0.21.1
prompt_toolkit-3.0.48 psutil-6.1.1 ptyprocess-0.7.0 pure-eval-0.2.3
pycparser-2.22 python-dateutil-2.9.0.post0 python-json-logger-3.2.1
pyyaml-6.0.2 pyzmq-26.2.0 referencing-0.35.1 rfc3339-validator-0.1.4
rfc3986-validator-0.1.1 rpds-py-0.22.3 send2trash-1.8.3 six-1.17.0
sniffio-1.3.1 soupsieve-2.6 stack_data-0.6.3 terminado-0.18.1
tinycss2-1.4.0 tornado-6.4.2 traitlets-5.14.3
types-python-dateutil-2.9.0.20241206 uri-template-1.3.0 wcwidth-0.2.13
webcolors-24.11.1 webencodings-0.5.1 websocket-client-1.8.0
$
```

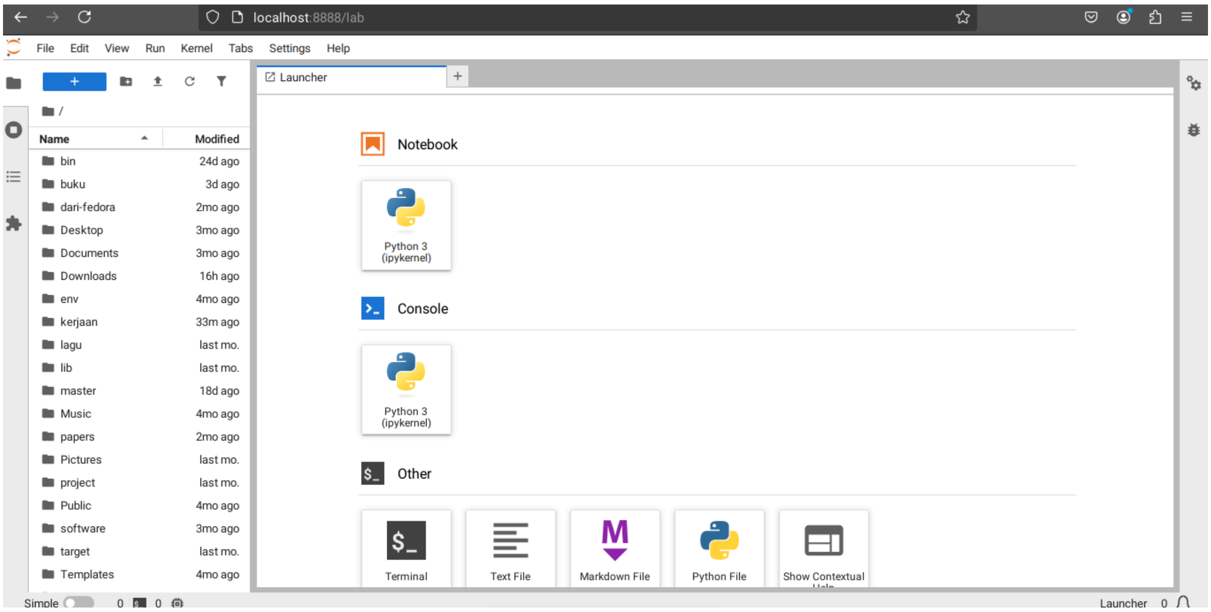
1.7 Run Jupyter Lab

```
$ jupyter lab
[I 2025-01-09 07:24:19.845 ServerApp] jupyter_lsp | extension was
successfully linked.
[I 2025-01-09 07:24:19.849 ServerApp] jupyter_server_terminals | extension
was successfully linked.
[I 2025-01-09 07:24:19.856 ServerApp] jupyterlab | extension was
successfully linked.
[I 2025-01-09 07:24:19.913 ServerApp] Writing Jupyter server cookie secret
to /home/bpdp/.local/share/jupyter/runtime/jupyter_cookie_secret
[I 2025-01-09 07:24:20.334 ServerApp] notebook_shim | extension was
successfully linked.
[I 2025-01-09 07:24:20.349 ServerApp] notebook_shim | extension was
successfully loaded.
[I 2025-01-09 07:24:20.351 ServerApp] jupyter_lsp | extension was
successfully loaded.
[I 2025-01-09 07:24:20.352 ServerApp] jupyter_server_terminals | extension
was successfully loaded.
[I 2025-01-09 07:24:20.388 LabApp] JupyterLab extension loaded from
/home/bpdp/software/python-dev-tools/micromamba-root/envs/py312-nlp/lib/pyt
hon3.12/site-packages/jupyterlab
```

```
[I 2025-01-09 07:24:20.389 LabApp] JupyterLab application directory is
/home/bpdp/software/python-dev-tools/micromamba-root/envs/py312-nlp/share/j
upyter/lab
[I 2025-01-09 07:24:20.390 LabApp] Extension Manager is 'pypi'.
[I 2025-01-09 07:24:20.443 ServerApp] jupyterlab | extension was
successfully loaded.
[I 2025-01-09 07:24:20.444 ServerApp] Serving notebooks from local
directory: /home/bpdp
[I 2025-01-09 07:24:20.444 ServerApp] Jupyter Server 2.15.0 is running at:
[I 2025-01-09 07:24:20.444 ServerApp]
http://localhost:8888/lab?token=85d4476fb60edebe645920a8c651f02724007e01067
d40b0
[I 2025-01-09 07:24:20.444 ServerApp]
http://127.0.0.1:8888/lab?token=85d4476fb60edebe645920a8c651f02724007e01067
d40b0
[I 2025-01-09 07:24:20.444 ServerApp] Use Control-C to stop this server and
shut down all kernels (twice to skip confirmation).
[C 2025-01-09 07:24:20.666 ServerApp]

    To access the server, open this file in a browser:

file:///home/bpdp/.local/share/jupyter/runtime/jpserver-13472-open.html
    Or copy and paste one of these URLs:

http://localhost:8888/lab?token=85d4476fb60edebe645920a8c651f02724007e01067
d40b0

http://127.0.0.1:8888/lab?token=85d4476fb60edebe645920a8c651f02724007e01067
d40b0
```

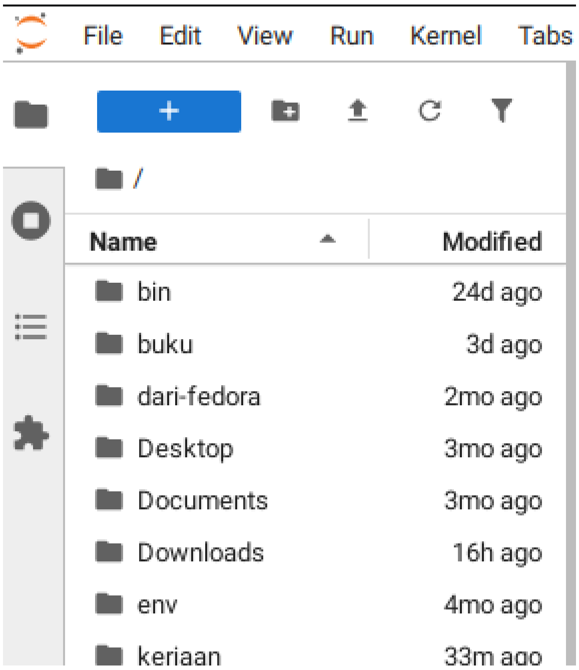
This will open the browser (new browser window if the browser is running, or open the browser if the browser is not running). See the log in your terminal if you want to see the URL, such as:

```
…
…
…
[C 2025-01-09 07:24:20.666 ServerApp]

    To access the server, open this file in a browser:

file:///home/bpdp/.local/share/jupyter/runtime/jpserver-13472-open.html
    Or copy and paste one of these URLs:

http://localhost:8888/lab?token=85d4476fb60edebe645920a8c651f02724007e01067
d40b0

http://127.0.0.1:8888/lab?token=85d4476fb60edebe645920a8c651f02724007e01067
d40b0
```

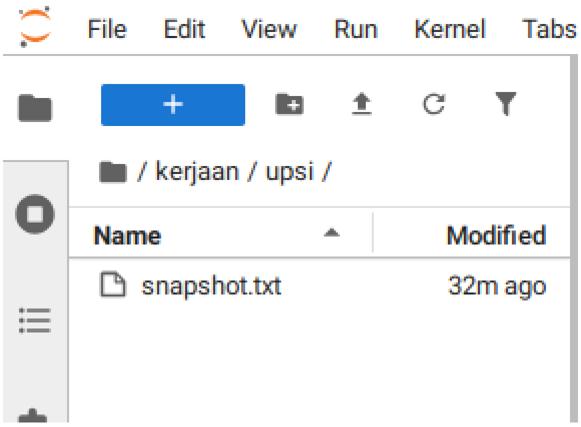


Name	Modified
bin	24d ago
buku	3d ago
dari-fedora	2mo ago
Desktop	3mo ago
Documents	3mo ago
Downloads	16h ago
env	4mo ago
kerjaan	33m ago
lagu	last mo.
lib	last mo.
master	18d ago
Music	4mo ago
papers	2mo ago
Pictures	last mo.
project	last mo.
Public	4mo ago
software	3mo ago
target	last mo.
Templates	4mo ago

On the left side, change the folder / directory to a directory which you want to use to save the workshop results (mine is `/kerjaan/upsi`):

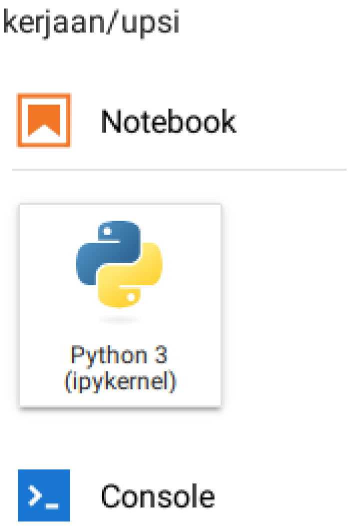


Name	Modified
bin	24d ago
buku	3d ago
dari-fedora	2mo ago
Desktop	3mo ago
Documents	3mo ago
Downloads	16h ago
env	4mo ago
kerjaan	33m ago

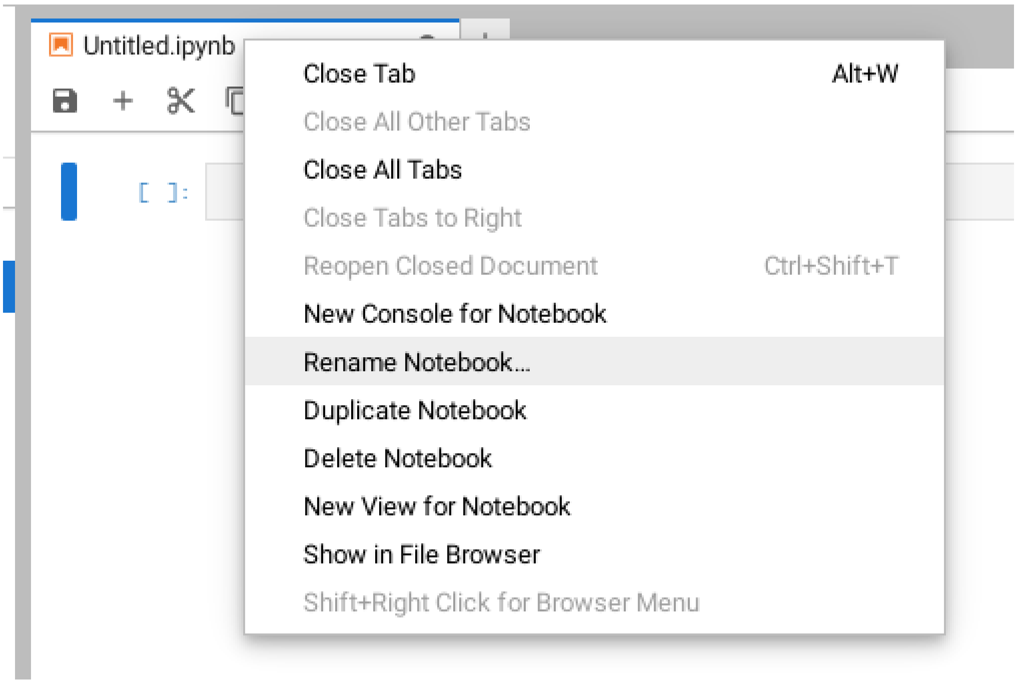


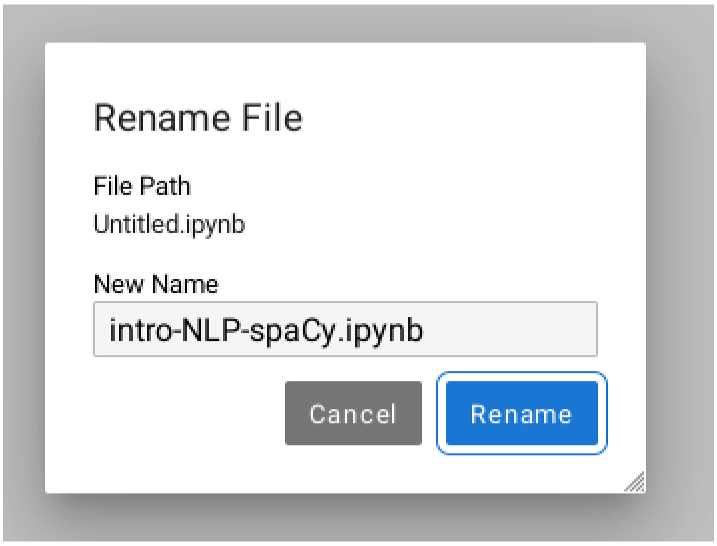
1.8 Create a Notebook (`.ipynb`)

Click on the ***`Python 3 (ipykernel) Console`*** to create a new notebook.



Jupyter will create a new notebook. Change the notebook name by using right-click on ***Untitled.ipynb*** name.



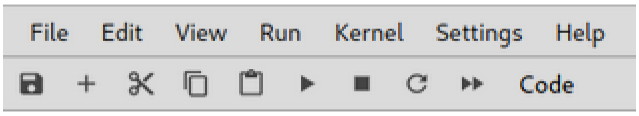

Rename File

File Path
Untitled.ipynb

New Name
intro-NLP-spaCy.ipynb

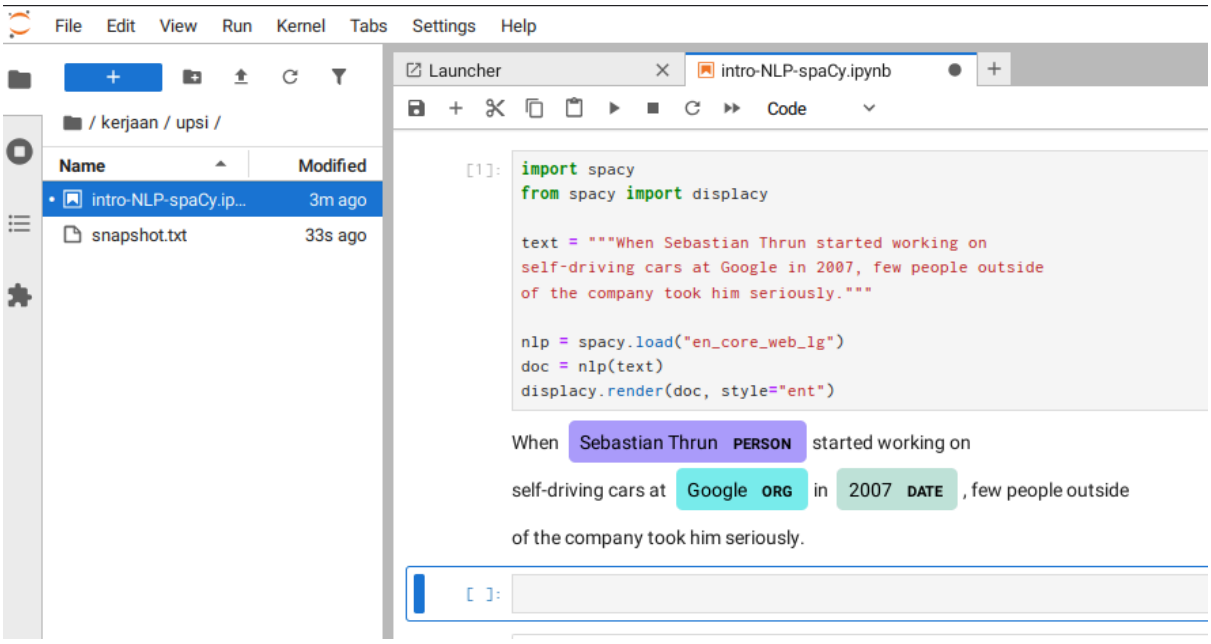
1.9 Try spaCy

To check that spaCy was installed properly, run this code in a cell (see the image below) by clicking on the cell and write the source code and run it using **Shift-Enter** or click the **Run** button on the menubar.



▶

```
import spacy
from spacy import displacy

text = """When Sebastian Thrun started working on
self-driving cars at Google in 2007, few people outside
of the company took him seriously."""

nlp = spacy.load("en_core_web_lg")
doc = nlp(text)
displacy.render(doc, style="ent")
```



```
import spacy
from spacy import displacy

text = """When Sebastian Thrun started working on
self-driving cars at Google in 2007, few people outside
of the company took him seriously."""

nlp = spacy.load("en_core_web_lg")
doc = nlp(text)
displacy.render(doc, style="ent")
```

When Sebastian Thrun PERSON started working on
self-driving cars at Google ORG in 2007 DATE , few people outside
of the company took him seriously.

If you have the results as shown above, congratulations, welcome to spaCy. Let’s explore NLP!