I081_Aniruddh_Kulkarni_NLP_Exp9

May 28, 2023

1 Name: Aniruddh Kulkarni

2 Roll no: I081

3 Stream: CS (AI)

4 Division: I

5 Semester: 5th Semester

6 Batch: I-3

7 Subject: NLP

8 Assignment-9

9 I) Web Scraping

[1]:
```
!pip install numpy==1.19.5
!pip install beautifulsoup4==4.6.3
```

```
Collecting numpy==1.19.5
  Using cached numpy-1.19.5.zip (7.3 MB)
  Installing build dependencies … done
  Getting requirements to build wheel … done
  Preparing metadata (pyproject.toml) … error
  error: subprocess-exited-with-error

  × Preparing metadata
(pyproject.toml) did not run successfully.
   exit code: 1
  > [255 lines of output]
      setup.py:67: RuntimeWarning: NumPy 1.19.5 may not yet support
Python 3.10.
        warnings.warn(
      Running from numpy source directory.
```

```
      setup.py:480: UserWarning: Unrecognized setuptools command,
proceeding with generating Cython sources and expanding templates
        run_build = parse_setuppy_commands()
      Processing numpy/random/_bounded_integers.pxd.in
      Processing numpy/random/_philox.pyx
      Processing numpy/random/_bounded_integers.pyx.in
      Processing numpy/random/_sfc64.pyx
      Processing numpy/random/_mt19937.pyx
      Processing numpy/random/bit_generator.pyx
      Processing numpy/random/mtrand.pyx
      Processing numpy/random/_generator.pyx
      Processing numpy/random/_pcg64.pyx
      Processing numpy/random/_common.pyx
      Cythonizing sources
      blas_opt_info:
      blas_mkl_info:
      customize UnixCCompiler
        libraries mkl_rt not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      blis_info:
        libraries blis not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      openblas_info:
      C compiler: clang -Wno-unused-result -Wsign-compare
-Wunreachable-code -DNDEBUG -fwrapv -O2 -Wall -fPIC -O2 -isystem
/Users/pushpakulkarni/miniconda3/include -arch arm64 -fPIC -O2 -isystem
/Users/pushpakulkarni/miniconda3/include -arch arm64

      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_/var
      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_/var/folders
      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_/var/folders/w7
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6
wsl_/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6
wsl_/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6
wsl_/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_
      compile options: '-c'
      clang:
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_/source.c
      clang /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl
```

```
_/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_/source.o
-L/Users/pushpakulkarni/miniconda3/lib -lopenblas -o
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6wsl_/a.out
      ld: warning: ignoring file
/Users/pushpakulkarni/miniconda3/lib/libopenblas.dylib, building for
macOS-x86_64 but attempting to link with file built for macOS-arm64
      ld: warning: ignoring file /var/folders/w7/x95pzq4s5s1f44mr9jmxg5
_m0000gn/T/tmpapo6wsl_/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpapo6ws
l_/source.o, building for macOS-x86_64 but attempting to link with file built
for unknown-arm64
      Undefined symbols for architecture x86_64:
        "_main", referenced from:
           implicit entry/start for main executable
      ld: symbol(s) not found for architecture x86_64
      clang: error: linker command failed with exit code 1 (use -v to
see invocation)
        NOT AVAILABLE

      atlas_3_10_blas_threads_info:
      Setting PTATLAS=ATLAS
        libraries tatlas not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      atlas_3_10_blas_info:
        libraries satlas not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      atlas_blas_threads_info:
      Setting PTATLAS=ATLAS
        libraries ptf77blas,ptcblas,atlas not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      atlas_blas_info:
        libraries f77blas,cblas,atlas not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      accelerate_info:
        libraries accelerate not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
      Library accelerate was not found. Ignoring
        libraries veclib not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
      Library veclib was not found. Ignoring
        FOUND:
```

```
          extra_compile_args = ['-msse3',
'-I/System/Library/Frameworks/vecLib.framework/Headers']
          extra_link_args = ['-Wl,-framework', '-Wl,Accelerate']
          define_macros = [('NO_ATLAS_INFO', 3), ('HAVE_CBLAS', None)]

        FOUND:
          extra_compile_args = ['-msse3',
'-I/System/Library/Frameworks/vecLib.framework/Headers']
          extra_link_args = ['-Wl,-framework', '-Wl,Accelerate']
          define_macros = [('NO_ATLAS_INFO', 3), ('HAVE_CBLAS', None)]

      non-existing path in 'numpy/distutils': 'site.cfg'
      lapack_opt_info:
      lapack_mkl_info:
        libraries mkl_rt not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      openblas_lapack_info:
      C compiler: clang -Wno-unused-result -Wsign-compare
-Wunreachable-code -DNDEBUG -fwrapv -O2 -Wall -fPIC -O2 -isystem
/Users/pushpakulkarni/miniconda3/include -arch arm64 -fPIC -O2 -isystem
/Users/pushpakulkarni/miniconda3/include -arch arm64

      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj/var
      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj/var/folders
      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj/var/folders/w7
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecy
sbsj/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecy
sbsj/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecy
sbsj/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj
      compile options: '-c'
      clang:
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj/source.c
      clang /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbs
j/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj/source.o
-L/Users/pushpakulkarni/miniconda3/lib -lopenblas -o
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysbsj/a.out
      ld: warning: ignoring file /var/folders/w7/x95pzq4s5s1f44mr9jmxg5
_m0000gn/T/tmpiecysbsj/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpiecysb
sj/source.o, building for macOS-x86_64 but attempting to link with file built
for unknown-arm64
      ld: warning: ignoring file
```

```
/Users/pushpakulkarni/miniconda3/lib/libopenblas.dylib, building for
macOS-x86_64 but attempting to link with file built for macOS-arm64
      Undefined symbols for architecture x86_64:
        "_main", referenced from:
           implicit entry/start for main executable
      ld: symbol(s) not found for architecture x86_64
      clang: error: linker command failed with exit code 1 (use -v to
see invocation)
        NOT AVAILABLE

      openblas_clapack_info:
      C compiler: clang -Wno-unused-result -Wsign-compare
-Wunreachable-code -DNDEBUG -fwrapv -O2 -Wall -fPIC -O2 -isystem
/Users/pushpakulkarni/miniconda3/include -arch arm64 -fPIC -O2 -isystem
/Users/pushpakulkarni/miniconda3/include -arch arm64

      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e/var
      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e/var/folders
      creating
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e/var/folders/w7
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7el
yo_e/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7el
yo_e/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T
      creating /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7el
yo_e/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e
      compile options: '-c'
      clang:
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e/source.c
      clang /var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_
e/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e/source.o
-L/Users/pushpakulkarni/miniconda3/lib -lopenblas -llapack -o
/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo_e/a.out
      ld: warning: ignoring file
/Users/pushpakulkarni/miniconda3/lib/liblapack.dylib, building for macOS-x86_64
but attempting to link with file built for macOS-arm64
      ld: warning: ignoring file /var/folders/w7/x95pzq4s5s1f44mr9jmxg5
_m0000gn/T/tmpf7elyo_e/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/tmpf7elyo
_e/source.o, building for macOS-x86_64 but attempting to link with file built
for unknown-arm64
      ld: warning: ignoring file
/Users/pushpakulkarni/miniconda3/lib/libopenblas.dylib, building for
macOS-x86_64 but attempting to link with file built for macOS-arm64
      Undefined symbols for architecture x86_64:
        "_main", referenced from:
           implicit entry/start for main executable
```

```
      ld: symbol(s) not found for architecture x86_64
      clang: error: linker command failed with exit code 1 (use -v to
see invocation)
        NOT AVAILABLE

      flame_info:
        libraries flame not found in
['/Users/pushpakulkarni/miniconda3/lib', '/usr/local/lib', '/usr/lib']
        NOT AVAILABLE

      atlas_3_10_threads_info:
      Setting PTATLAS=ATLAS
        libraries lapack_atlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries tatlas,tatlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries lapack_atlas not found in /usr/local/lib
        libraries tatlas,tatlas not found in /usr/local/lib
        libraries lapack_atlas not found in /usr/lib
        libraries tatlas,tatlas not found in /usr/lib
      <class 'numpy.distutils.system_info.atlas_3_10_threads_info'>
        NOT AVAILABLE

      atlas_3_10_info:
        libraries lapack_atlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries satlas,satlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries lapack_atlas not found in /usr/local/lib
        libraries satlas,satlas not found in /usr/local/lib
        libraries lapack_atlas not found in /usr/lib
        libraries satlas,satlas not found in /usr/lib
      <class 'numpy.distutils.system_info.atlas_3_10_info'>
        NOT AVAILABLE

      atlas_threads_info:
      Setting PTATLAS=ATLAS
        libraries lapack_atlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries ptf77blas,ptcblas,atlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries lapack_atlas not found in /usr/local/lib
        libraries ptf77blas,ptcblas,atlas not found in /usr/local/lib
        libraries lapack_atlas not found in /usr/lib
        libraries ptf77blas,ptcblas,atlas not found in /usr/lib
      <class 'numpy.distutils.system_info.atlas_threads_info'>
        NOT AVAILABLE
```

```
      atlas_info:
        libraries lapack_atlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries f77blas,cblas,atlas not found in
/Users/pushpakulkarni/miniconda3/lib
        libraries lapack_atlas not found in /usr/local/lib
        libraries f77blas,cblas,atlas not found in /usr/local/lib
        libraries lapack_atlas not found in /usr/lib
        libraries f77blas,cblas,atlas not found in /usr/lib
      <class 'numpy.distutils.system_info.atlas_info'>
        NOT AVAILABLE

        FOUND:
          extra_compile_args = ['-msse3',
'-I/System/Library/Frameworks/vecLib.framework/Headers']
          extra_link_args = ['-Wl,-framework', '-Wl,Accelerate']
          define_macros = [('NO_ATLAS_INFO', 3), ('HAVE_CBLAS', None)]

      /private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-
build-env-alq35yrg/overlay/lib/python3.10/site-
packages/setuptools/_distutils/dist.py:275: UserWarning: Unknown distribution
option: 'define_macros'
        warnings.warn(msg)
      running dist_info
      running build_src
      build_src
      building py_modules sources
      creating build
      creating build/src.macosx-11.1-arm64-3.10
      creating build/src.macosx-11.1-arm64-3.10/numpy
      creating build/src.macosx-11.1-arm64-3.10/numpy/distutils
      building library "npymath" sources
      Could not locate executable gfortran
      Could not locate executable f95
      Could not locate executable f90
      Could not locate executable f77
      Could not locate executable xlf90
      Could not locate executable xlf
      Could not locate executable ifort
      Could not locate executable ifc
      Could not locate executable g77
      Could not locate executable g95
      Could not locate executable pgfortran
      don't know how to compile Fortran code on platform 'posix'
      ld: warning: ignoring file _configtest.o, building for
macOS-x86_64 but attempting to link with file built for unknown-arm64
      Undefined symbols for architecture x86_64:
        "_main", referenced from:
```

```
           implicit entry/start for main executable
      ld: symbol(s) not found for architecture x86_64
      clang: error: linker command failed with exit code 1 (use -v to
see invocation)
      Traceback (most recent call last):
        File "/Users/pushpakulkarni/miniconda3/lib/python3.10/site-
packages/pip/_vendor/pyproject_hooks/_in_process/_in_process.py", line 353, in
<module>
          main()
        File "/Users/pushpakulkarni/miniconda3/lib/python3.10/site-
packages/pip/_vendor/pyproject_hooks/_in_process/_in_process.py", line 335, in
main
          json_out['return_val'] = hook(**hook_input['kwargs'])
        File "/Users/pushpakulkarni/miniconda3/lib/python3.10/site-
packages/pip/_vendor/pyproject_hooks/_in_process/_in_process.py", line 149, in
prepare_metadata_for_build_wheel
          return hook(metadata_directory, config_settings)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/build_meta.py", line
157, in prepare_metadata_for_build_wheel
          self.run_setup()
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/build_meta.py", line
248, in run_setup
          super(_BuildMetaLegacyBackend,
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/build_meta.py", line
142, in run_setup
          exec(compile(code, __file__, 'exec'), locals())
        File "setup.py", line 508, in <module>
          setup_package()
        File "setup.py", line 500, in setup_package
          setup(**metadata)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-install-38e1rjd6/n
umpy_5856890b9140438ab30afbde04f55194/numpy/distutils/core.py", line 169, in
setup
          return old_setup(**new_attr)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/__init__.py", line 165,
in setup
          return distutils.core.setup(**attrs)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
```

```
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/_distutils/core.py",
line 148, in setup
          dist.run_commands()
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/_distutils/dist.py",
line 967, in run_commands
          self.run_command(cmd)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/_distutils/dist.py",
line 986, in run_command
          cmd_obj.run()
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/command/dist_info.py",
line 31, in run
          egg_info.run()
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-install-38e1rjd6/n
umpy_5856890b9140438ab30afbde04f55194/numpy/distutils/command/egg_info.py", line
24, in run
          self.run_command("build_src")
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/_distutils/cmd.py",
line 313, in run_command
          self.distribution.run_command(command)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-build-env-
alq35yrg/overlay/lib/python3.10/site-packages/setuptools/_distutils/dist.py",
line 986, in run_command
          cmd_obj.run()
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-install-38e1rjd6/n
umpy_5856890b9140438ab30afbde04f55194/numpy/distutils/command/build_src.py",
line 144, in run
          self.build_sources()
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-install-38e1rjd6/n
umpy_5856890b9140438ab30afbde04f55194/numpy/distutils/command/build_src.py",
line 155, in build_sources
          self.build_library_sources(*libname_info)
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-install-38e1rjd6/n
umpy_5856890b9140438ab30afbde04f55194/numpy/distutils/command/build_src.py",
line 288, in build_library_sources
          sources = self.generate_sources(sources, (lib_name,
```

```
build_info))
        File
"/private/var/folders/w7/x95pzq4s5s1f44mr9jmxg5_m0000gn/T/pip-install-38e1rjd6/n
umpy_5856890b9140438ab30afbde04f55194/numpy/distutils/command/build_src.py",
line 378, in generate_sources
          source = func(extension, build_dir)
        File "numpy/core/setup.py", line 663, in get_mathlib_info
          raise RuntimeError("Broken toolchain: cannot link a simple C
program")
      RuntimeError: Broken toolchain: cannot link a simple C program
      [end of output]

  note: This error originates from a subprocess, and is likely not a
problem with pip.
error: metadata-generation-failed

× Encountered error while generating package metadata.
 > See above for output.

note: This is an issue with the package mentioned above, not pip.
hint: See above for details.
Requirement already satisfied: beautifulsoup4==4.6.3 in
/Users/pushpakulkarni/miniconda3/lib/python3.10/site-packages (4.6.3)
```

[2]:
```
# making the necessary imports
from pprint import pprint
from bs4 import BeautifulSoup
from urllib.request import urlopen
```

[3]:
```
myurl = "https://old.reddit.com/r/datascience" # specify the url
html = urlopen(myurl).read() # query the website so that it returns a html page␣
 ↪
soupified = BeautifulSoup(html, 'html.parser') # parse the html in the 'html'␣
 ↪variable, and store it in Beautiful Soup format
```

[4]:
```
pprint(soupified.prettify()[:2000]) # to get an idea of the html structure of␣
 ↪the webpage
```

```
('<!DOCTYPE html>\n'
 '<html lang="en" xml:lang="en" xmlns="http://www.w3.org/1999/xhtml">\n'
 ' <head>\n'
 '  <title>\n'
 '   Data Science\n'
 '  </title>\n'
 '  <meta content=" reddit, reddit.com, vote, comment, submit " '
 'name="keywords"/>\n'
 '  <meta content="A place for data science practitioners and professionals to '
 'discuss and debate data science career questions." name="description"/>\n'
```

```
 '  <meta content="always" name="referrer"/>\n'
 '  <meta content="text/html; charset=utf-8" http-equiv="Content-Type">\n'
 '   <link href="/static/opensearch.xml" rel="search" '
 'type="application/opensearchdescription+xml"/>\n'
 '   <link href="https://www.reddit.com/r/datascience/" rel="canonical"/>\n'
 '   <meta content="width=1024" name="viewport"/>\n'
 '   <link href="//out.reddit.com" rel="dns-prefetch"/>\n'
 '   <link href="//out.reddit.com" rel="preconnect"/>\n'
 '   <meta '
 'content="https://styles.redditmedia.com/t5_2sptq/styles/communityIcon_fdj8zurr
ifa71.png" '
 'property="og:image"/>\n'
 '   <meta content="reddit" property="og:site_name"/>\n'
 '   <meta content="A place for data science practitioners and professionals '
 'to discuss and debate data science career questions." '
 'property="og:description"/>\n'
 '   <meta content="Data Science • r/datascience" property="og:title"/>\n'
 '   <meta content="com.reddit.frontpage" property="al:android:package"/>\n'
 '   <meta content="Reddit" property="al:ios:app_name"/>\n'
 '   <meta content="reddit://www.reddit.com/r/datascience/" '
 'property="al:ios:url"/>\n'
 '   <meta content="1064216828" property="al:ios:app_store_id"/>\n'
 '   <meta content="reddit" property="twitter:site"/>\n'
 '   <meta content="summary" property="twitter:card"/>\n'
 '   <meta content="Data Science • r/datascience" property="twitter:title"/>\n'
 '   <link '
 'href="//www.redditstatic.com/desktop2x/img/favicon/apple-icon-57x57.png" '
 'rel="apple-touch-icon" sizes="57x57">\n'
 '    <link '
 'href="//www.redditstatic.com/desktop2x/img/favicon/apple-icon-60x60.png" '
 'rel="apple-touch-icon" sizes="60x60">\n'
 '     <link '
 'href="//www.redditstatic.com/desktop2x/img/favicon/apple-icon-72x72.png" '
 'rel="apple-touch-icon" sizes')
```

[5]:
```
soupified.title # to get the title of the web page
```

[5]:
```
<title>Data Science</title>
```

[6]:
```
import requests
# Headers to mimic a browser visit
headers = {'User-Agent': 'Mozilla/5.0'}

# Returns a requests.models.Response object
page = requests.get(myurl, headers=headers)
```

[7]:
```
soupified = BeautifulSoup(page.text, 'html.parser')
```

[8]:
```
domains = soupified.find_all("span", class_="domain")  #looks into itself for␣
↪all the anchor tags with the parameters passed in as the second argument.
```

[9]:
```
global domain
for domain in domains:
    if domain != "(self.datascience)":
        continue

    print(domain.text)
```

[10]:
```
global parent_div
for domain in soupified.find_all("span", class_="domain"):
    if domain != "(self.datascience)":
        continue

    parent_div = domain.parent.parent.parent.parent
    print(parent_div.text)
```

[11]:
```
attrs = {'class': 'thing', 'data-domain': 'self.datascience'}

for post in soupified.find_all('div', attrs=attrs):
    print(post.attrs['data-domain'])
```

```
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
self.datascience
```

[12]:
```
title = post.find('p', class_="title").text
title
```

[12]:
```
"CareerRecent Graduate's Offer was Rescinded (self.datascience)"
```

[13]:
```
author = post.find('a', class_='author').text
author
```

[13]:
```
'wavehnter'
```

[14]:
```
comments = post.find('a', class_='comments').text
comments
```

[14]:
```
'7 comments'
```

[15]:
```
likes = post.find("div", attrs={"class": "score likes"}).text
likes
```

[15]:
```
'0'
```

[17]:
```
import csv
counter = 1
for post in soupified.find_all('div', attrs=attrs):
    post_line = [counter, title, author, likes, comments]
    with open('output.csv', 'a') as f:
        writer = csv.writer(f)
        writer.writerow(post_line)

    counter += 1
```

[18]:
```
counter
```

[18]:
```
21
```

[19]:
```
import pandas as pd
a = pd.read_csv("output.csv" )
```

[20]:
```
a.head
```

[20]:
```
<bound method NDFrame.head of      1 CareerRecent Graduate's Offer was Rescinded
(self.datascience)
0    2  CareerRecent Graduate's Offer was Rescinded (s…              \
1    3  CareerRecent Graduate's Offer was Rescinded (s…
2    4  CareerRecent Graduate's Offer was Rescinded (s…
3    5  CareerRecent Graduate's Offer was Rescinded (s…
4    6  CareerRecent Graduate's Offer was Rescinded (s…
5    7  CareerRecent Graduate's Offer was Rescinded (s…
6    8  CareerRecent Graduate's Offer was Rescinded (s…
7    9  CareerRecent Graduate's Offer was Rescinded (s…
8   10  CareerRecent Graduate's Offer was Rescinded (s…
```

```
9   11  CareerRecent Graduate's Offer was Rescinded (s…
10  12  CareerRecent Graduate's Offer was Rescinded (s…
11  13  CareerRecent Graduate's Offer was Rescinded (s…
12  14  CareerRecent Graduate's Offer was Rescinded (s…
13  15  CareerRecent Graduate's Offer was Rescinded (s…
14  16  CareerRecent Graduate's Offer was Rescinded (s…
15  17  CareerRecent Graduate's Offer was Rescinded (s…
16  18  CareerRecent Graduate's Offer was Rescinded (s…
17  19  CareerRecent Graduate's Offer was Rescinded (s…
18  20  CareerRecent Graduate's Offer was Rescinded (s…

    wavehnter  0  7 comments
0   wavehnter  0  7 comments
1   wavehnter  0  7 comments
2   wavehnter  0  7 comments
3   wavehnter  0  7 comments
4   wavehnter  0  7 comments
5   wavehnter  0  7 comments
6   wavehnter  0  7 comments
7   wavehnter  0  7 comments
8   wavehnter  0  7 comments
9   wavehnter  0  7 comments
10  wavehnter  0  7 comments
11  wavehnter  0  7 comments
12  wavehnter  0  7 comments
13  wavehnter  0  7 comments
14  wavehnter  0  7 comments
15  wavehnter  0  7 comments
16  wavehnter  0  7 comments
17  wavehnter  0  7 comments
18  wavehnter  0  7 comments  >
```

[21]:
```
a.shape
```

[21]:
```
(19, 5)
```

10 II) Regex

[22]:
```
corpus = "User: I am unhappy, ahhhhh....huhhh...mehhh...hhhh. \
ELIZA: DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY?\
User: I need some help!!, that much seems certain. Can pay  2 Million $bucks␣
 ↪for getting happy.\
ELIZA: WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP\
User: Perhaps I could learn to get along with my mother.\
ELIZA: TELL ME MORE ABOUT YOUR FAMILY \
User: My mother takes care of me.\
```

```
ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU? \
User: My father.\
ELIZA: YOUR FATHER \
User: You are like my father in some ways. \
ELIZA: You can always talk to me and also access my program on stanford.edu and␣
 ↪can call for enquiry on +1-420-8374269"


#The dialogue above is from ELIZA, an early natural language processing system
#that could carry on a limited conversation with a user by imitating the␣
 ↪responses of a Rogerian psychotherapist
```

[23]:
```
import re
corpus
```

[23]:
```
'User: I am unhappy, ahhhhh…huhhh…mehhh…hhhh. ELIZA: DO YOU THINK COMING
HERE WILL HELP YOU NOT TO BE UNHAPPY?User: I need some help!!, that much seems
certain. Can pay  2 Million $bucks for getting happy.ELIZA: WHAT WOULD IT MEAN
TO YOU IF YOU GOT SOME HELPUser: Perhaps I could learn to get along with my
mother.ELIZA: TELL ME MORE ABOUT YOUR FAMILY User: My mother takes care of
me.ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU? User: My father.ELIZA: YOUR
FATHER User: You are like my father in some ways. ELIZA: You can always talk to
me and also access my program on stanford.edu and can call for enquiry on
+1-420-8374269'
```

[24]:
```
substituted_string = re.sub(r'\S+(.edu){1}', '', corpus) #remove website from␣
 ↪the given text.
print(substituted_string)
```

```
User: I am unhappy, ahhhhh…huhhh…mehhh…hhhh. ELIZA: DO YOU THINK COMING
HERE WILL HELP YOU NOT TO BE UNHAPPY?User: I need some help!!, that much seems
certain. Can pay  2 Million $bucks for getting happy.ELIZA: WHAT WOULD IT MEAN
TO YOU IF YOU GOT SOME HELPUser: Perhaps I could learn to get along with my
mother.ELIZA: TELL ME MORE ABOUT YOUR FAMILY User: My mother takes care of
me.ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU? User: My father.ELIZA: YOUR
FATHER User: You are like my father in some ways. ELIZA: You can always talk to
me and also access my program on  and can call for enquiry on +1-420-8374269
```

[25]:
```
re.findall('\S+[.com]{1}|\+[0-9]{1}-[0-9]{3}-[0-9]{7}', corpus)
#tagging all contact information as hidden
redacted = re.sub('\S+[.edu]{1}|\+[0-9]{1}-[0-9]{3}-[0-9]{7}', '<HIDDEN>',␣
 ↪corpus)

print(redacted)
```

```
<HIDDEN>r: I am unhappy, <HIDDEN> ELIZA: DO YOU THINK COMING HERE WILL HELP YOU
NOT TO BE <HIDDEN>r: I <HIDDEN> <HIDDEN> <HIDDEN>lp!!, that <HIDDEN>ch
```

<HIDDEN>ms <HIDDEN> Can pay 2 Million <HIDDEN>cks for <HIDDEN>tting
<HIDDEN>ELIZA: WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME <HIDDEN>r:
<HIDDEN>rhaps I <HIDDEN> <HIDDEN>arn to <HIDDEN>t along with my <HIDDEN>ELIZA:
TELL ME MORE ABOUT YOUR FAMILY <HIDDEN>r: My <HIDDEN>r <HIDDEN>s <HIDDEN> of
<HIDDEN>ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU? <HIDDEN>r: My
<HIDDEN>ELIZA: YOUR FATHER <HIDDEN>r: <HIDDEN> <HIDDEN> <HIDDEN> my <HIDDEN>r in
<HIDDEN> <HIDDEN> ELIZA: <HIDDEN> can always talk to <HIDDEN> <HIDDEN> also
<HIDDEN>ss my program on <HIDDEN> <HIDDEN> can call for <HIDDEN>iry on <HIDDEN>

[26]:
```
#Check if the string contains any digits (numbers from 0-9):

x = re.findall("\d", corpus)
print(x)

if (x):
  print("Yes, there is at least one match!")
else:
  print("No match")
```

```
['2', '1', '4', '2', '0', '8', '3', '7', '4', '2', '6', '9']
Yes, there is at least one match!
```

[27]:
```
def multi_re_find(patterns,phrase):
    '''
    Takes in a list of regex patterns
    Prints a list of all matches
    '''
    for pattern in patterns:
        print ('Searching the phrase using the re check: %r' %pattern)
        print (re.findall(pattern,phrase))
        print ('\n')
```

[28]:
```
test_patterns=[ '[a-z]+',      # sequences of lower case letters
                '[A-Z]+',      # sequences of upper case letters
                '[a-zA-Z]+',   # sequences of lower or upper case letters
                '[A-Z][a-z]+'] # one upper case letter followed by lower case␣
 ↪letters

multi_re_find(test_patterns,corpus)
```

```
Searching the phrase using the re check: '[a-z]+'
['ser', 'am', 'unhappy', 'ahhhhh', 'huhhh', 'mehhh', 'hhhh', 'ser', 'need',
'some', 'help', 'that', 'much', 'seems', 'certain', 'an', 'pay', 'illion',
'bucks', 'for', 'getting', 'happy', 'ser', 'erhaps', 'could', 'learn', 'to',
'get', 'along', 'with', 'my', 'mother', 'ser', 'y', 'mother', 'takes', 'care',
'of', 'me', 'ser', 'y', 'father', 'ser', 'ou', 'are', 'like', 'my', 'father',
'in', 'some', 'ways', 'ou', 'can', 'always', 'talk', 'to', 'me', 'and', 'also',
'access', 'my', 'program', 'on', 'stanford', 'edu', 'and', 'can', 'call', 'for',
```

```
'enquiry', 'on']


Searching the phrase using the re check: '[A-Z]+'
['U', 'I', 'ELIZA', 'DO', 'YOU', 'THINK', 'COMING', 'HERE', 'WILL', 'HELP',
'YOU', 'NOT', 'TO', 'BE', 'UNHAPPY', 'U', 'I', 'C', 'M', 'ELIZA', 'WHAT',
'WOULD', 'IT', 'MEAN', 'TO', 'YOU', 'IF', 'YOU', 'GOT', 'SOME', 'HELPU', 'P',
'I', 'ELIZA', 'TELL', 'ME', 'MORE', 'ABOUT', 'YOUR', 'FAMILY', 'U', 'M',
'ELIZA', 'WHO', 'ELSE', 'IN', 'YOU', 'FAMILY', 'TAKES', 'CARE', 'OF', 'YOU',
'U', 'M', 'ELIZA', 'YOUR', 'FATHER', 'U', 'Y', 'ELIZA', 'Y']


Searching the phrase using the re check: '[a-zA-Z]+'
['User', 'I', 'am', 'unhappy', 'ahhhhh', 'huhhh', 'mehhh', 'hhhh', 'ELIZA',
'DO', 'YOU', 'THINK', 'COMING', 'HERE', 'WILL', 'HELP', 'YOU', 'NOT', 'TO',
'BE', 'UNHAPPY', 'User', 'I', 'need', 'some', 'help', 'that', 'much', 'seems',
'certain', 'Can', 'pay', 'Million', 'bucks', 'for', 'getting', 'happy', 'ELIZA',
'WHAT', 'WOULD', 'IT', 'MEAN', 'TO', 'YOU', 'IF', 'YOU', 'GOT', 'SOME',
'HELPUser', 'Perhaps', 'I', 'could', 'learn', 'to', 'get', 'along', 'with',
'my', 'mother', 'ELIZA', 'TELL', 'ME', 'MORE', 'ABOUT', 'YOUR', 'FAMILY',
'User', 'My', 'mother', 'takes', 'care', 'of', 'me', 'ELIZA', 'WHO', 'ELSE',
'IN', 'YOU', 'FAMILY', 'TAKES', 'CARE', 'OF', 'YOU', 'User', 'My', 'father',
'ELIZA', 'YOUR', 'FATHER', 'User', 'You', 'are', 'like', 'my', 'father', 'in',
'some', 'ways', 'ELIZA', 'You', 'can', 'always', 'talk', 'to', 'me', 'and',
'also', 'access', 'my', 'program', 'on', 'stanford', 'edu', 'and', 'can',
'call', 'for', 'enquiry', 'on']


Searching the phrase using the re check: '[A-Z][a-z]+'
['User', 'User', 'Can', 'Million', 'User', 'Perhaps', 'User', 'My', 'User',
'My', 'User', 'You', 'You']
```

[29]:
```
test_patterns = [ 'ah*',     # a followed by zero or more h's
                  'EL+',        # E followed by one or more L's
                  'me?',        # m followed by zero or one e's
                  'huh{2}',      # h followed by two h's
                  'hh{2,3}',    # h followed by two to three h's
                 ]

multi_re_find(test_patterns,corpus)
```

```
Searching the phrase using the re check: 'ah*'
['a', 'a', 'ahhhhh', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a',
'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a']
```

```
Searching the phrase using the re check: 'EL+'
['EL', 'EL', 'EL', 'EL', 'EL', 'ELL', 'EL', 'EL', 'EL', 'EL']


Searching the phrase using the re check: 'me?'
['m', 'me', 'me', 'm', 'm', 'm', 'm', 'm', 'me', 'm', 'me', 'me', 'm', 'm']


Searching the phrase using the re check: 'huh{2}'
['huhh']


Searching the phrase using the re check: 'hh{2,3}'
['hhhh', 'hhh', 'hhh', 'hhhh']
```

[30]:
```
test_patterns=[ r'\d+', # sequence of digits
                r'\D+', # sequence of non-digits
                r'\s+', # sequence of whitespace
                r'\S+', # sequence of non-whitespace
                r'\w+', # alphanumeric characters
                r'\W+', # non-alphanumeric
                ]

multi_re_find(test_patterns,corpus)
```

```
Searching the phrase using the re check: '\\d+'
['2', '1', '420', '8374269']


Searching the phrase using the re check: '\\D+'
['User: I am unhappy, ahhhhh…huhhh…mehhh…hhhh. ELIZA: DO YOU THINK COMING
HERE WILL HELP YOU NOT TO BE UNHAPPY?User: I need some help!!, that much seems
certain. Can pay  ', ' Million $bucks for getting happy.ELIZA: WHAT WOULD IT
MEAN TO YOU IF YOU GOT SOME HELPUser: Perhaps I could learn to get along with my
mother.ELIZA: TELL ME MORE ABOUT YOUR FAMILY User: My mother takes care of
me.ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU? User: My father.ELIZA: YOUR
FATHER User: You are like my father in some ways. ELIZA: You can always talk to
me and also access my program on stanford.edu and can call for enquiry on +',
'-', '-']


Searching the phrase using the re check: '\\s+'
[' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', '  ', ' ', ' ', ' ', ' ',
' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
```

```
' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ']


Searching the phrase using the re check: '\\S+'
['User:', 'I', 'am', 'unhappy,', 'ahhhhh…huhhh…mehhh…hhhh.', 'ELIZA:',
'DO', 'YOU', 'THINK', 'COMING', 'HERE', 'WILL', 'HELP', 'YOU', 'NOT', 'TO',
'BE', 'UNHAPPY?User:', 'I', 'need', 'some', 'help!!,', 'that', 'much', 'seems',
'certain.', 'Can', 'pay', '2', 'Million', '$bucks', 'for', 'getting',
'happy.ELIZA:', 'WHAT', 'WOULD', 'IT', 'MEAN', 'TO', 'YOU', 'IF', 'YOU', 'GOT',
'SOME', 'HELPUser:', 'Perhaps', 'I', 'could', 'learn', 'to', 'get', 'along',
'with', 'my', 'mother.ELIZA:', 'TELL', 'ME', 'MORE', 'ABOUT', 'YOUR', 'FAMILY',
'User:', 'My', 'mother', 'takes', 'care', 'of', 'me.ELIZA:', 'WHO', 'ELSE',
'IN', 'YOU', 'FAMILY', 'TAKES', 'CARE', 'OF', 'YOU?', 'User:', 'My',
'father.ELIZA:', 'YOUR', 'FATHER', 'User:', 'You', 'are', 'like', 'my',
'father', 'in', 'some', 'ways.', 'ELIZA:', 'You', 'can', 'always', 'talk', 'to',
'me', 'and', 'also', 'access', 'my', 'program', 'on', 'stanford.edu', 'and',
'can', 'call', 'for', 'enquiry', 'on', '+1-420-8374269']


Searching the phrase using the re check: '\\w+'
['User', 'I', 'am', 'unhappy', 'ahhhhh', 'huhhh', 'mehhh', 'hhhh', 'ELIZA',
'DO', 'YOU', 'THINK', 'COMING', 'HERE', 'WILL', 'HELP', 'YOU', 'NOT', 'TO',
'BE', 'UNHAPPY', 'User', 'I', 'need', 'some', 'help', 'that', 'much', 'seems',
'certain', 'Can', 'pay', '2', 'Million', 'bucks', 'for', 'getting', 'happy',
'ELIZA', 'WHAT', 'WOULD', 'IT', 'MEAN', 'TO', 'YOU', 'IF', 'YOU', 'GOT', 'SOME',
'HELPUser', 'Perhaps', 'I', 'could', 'learn', 'to', 'get', 'along', 'with',
'my', 'mother', 'ELIZA', 'TELL', 'ME', 'MORE', 'ABOUT', 'YOUR', 'FAMILY',
'User', 'My', 'mother', 'takes', 'care', 'of', 'me', 'ELIZA', 'WHO', 'ELSE',
'IN', 'YOU', 'FAMILY', 'TAKES', 'CARE', 'OF', 'YOU', 'User', 'My', 'father',
'ELIZA', 'YOUR', 'FATHER', 'User', 'You', 'are', 'like', 'my', 'father', 'in',
'some', 'ways', 'ELIZA', 'You', 'can', 'always', 'talk', 'to', 'me', 'and',
'also', 'access', 'my', 'program', 'on', 'stanford', 'edu', 'and', 'can',
'call', 'for', 'enquiry', 'on', '1', '420', '8374269']


Searching the phrase using the re check: '\\W+'
[': ', ' ', ' ', ', ', '…', '…', '…', '. ', ': ', ' ', ' ', ' ', ' ', '
', ' ', ' ', ' ', ' ', ' ', ' ', ' ', '?', ': ', ' ', ' ', ' ', '!!, ', ' ', ' ', '
', '. ', ' ', '  ', ' ', ' $', ' ', ' ', ' ', '.', ': ', ' ', ' ', ' ', ' ', '
', ' ', ' ', ' ', ' ', ' ', ' ', ': ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
'.', ': ', ' ', ' ', ' ', ' ', ' ', ' ', ': ', ' ', ' ', ' ', ' ', ' ', '.', ':
', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', '? ', ': ', ' ', '.', ': ', ' ', ' ',
': ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', '. ', ': ', ' ', ' ', ' ', ' ', ' ', '
', ' ', ' ', ' ', ' ', ' ', ' ', '.', ' ', ' ', ' ', ' ', ' ', ' ', ' +', '-',
'-']
```

[31]:
```
re.findall('[^!.? ]+',corpus)
```

[31]:
```
['User:',
 'I',
 'am',
 'unhappy,',
 'ahhhhh',
 'huhhh',
 'mehhh',
 'hhhh',
 'ELIZA:',
 'DO',
 'YOU',
 'THINK',
 'COMING',
 'HERE',
 'WILL',
 'HELP',
 'YOU',
 'NOT',
 'TO',
 'BE',
 'UNHAPPY',
 'User:',
 'I',
 'need',
 'some',
 'help',
 ',',
 'that',
 'much',
 'seems',
 'certain',
 'Can',
 'pay',
 '2',
 'Million',
 '$bucks',
 'for',
 'getting',
 'happy',
 'ELIZA:',
 'WHAT',
 'WOULD',
 'IT',
```

```
'MEAN',
'TO',
'YOU',
'IF',
'YOU',
'GOT',
'SOME',
'HELPUser:',
'Perhaps',
'I',
'could',
'learn',
'to',
'get',
'along',
'with',
'my',
'mother',
'ELIZA:',
'TELL',
'ME',
'MORE',
'ABOUT',
'YOUR',
'FAMILY',
'User:',
'My',
'mother',
'takes',
'care',
'of',
'me',
'ELIZA:',
'WHO',
'ELSE',
'IN',
'YOU',
'FAMILY',
'TAKES',
'CARE',
'OF',
'YOU',
'User:',
'My',
'father',
'ELIZA:',
'YOUR',
```

```
'FATHER',
'User:',
'You',
'are',
'like',
'my',
'father',
'in',
'some',
'ways',
'ELIZA:',
'You',
'can',
'always',
'talk',
'to',
'me',
'and',
'also',
'access',
'my',
'program',
'on',
'stanford',
'edu',
'and',
'can',
'call',
'for',
'enquiry',
'on',
'+1-420-8374269']
```

11 Conclusion

1) Anything which we can see on the web can be scraped and stored locally. We can use the information we just acquired for a multitude of purposes. With just the fours pieces of information we found, we can draw up a lot of conclusions.

2) By analyzing our information further, we can figure which kind of posts receive the most likes, what words they contain, who they're posted by. Or conversely, we can find make a controversy calculator by analyzing the ratio of likes to comments.

3) BeautifulSoup is one of the many libraries which allow us to scrape web pages. Depending on our needs we can choose between the many available choices like beautifulsoup, scrapy, selenium, etc.

4) It's important to keep in mind that it's pretty nice of the website even to allow us to scrape them because if they wanted, they could detect bots in the first 10 to 20 requests, or even catch us based on the request object that Python sends. So its our responsibility to Scrape responsibly and not overload the host server.

5) A regular expression (shortened as regex or regexp; sometimes referred to as rational expression) is a sequence of characters that specifies a search pattern in text. Usually such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation.

6) 're' module provides regular expression matching operations similar to those found in Perl.

7) Both patterns and strings to be searched can be Unicode strings (str) as well as 8-bit strings (bytes). However, Unicode strings and 8-bit strings cannot be mixed