

자연어처리 프로젝트 2차발표

2조

2013210043 권민규

2014210035 전수혁

2014210064 변지석

2019/May/22

Python Document Search

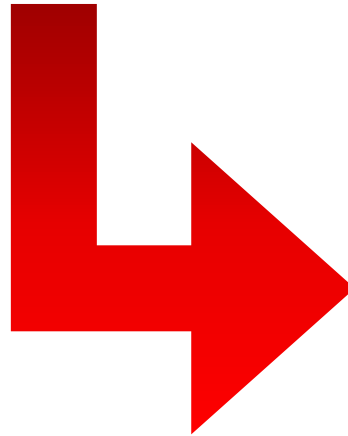
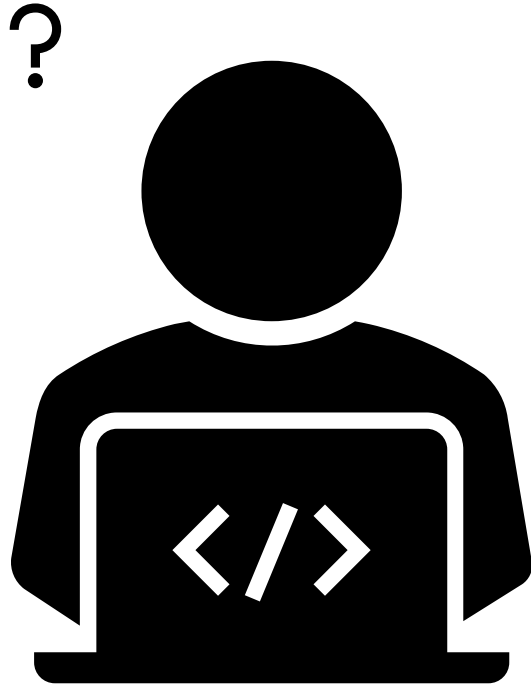
파이썬 문서 검색기

Why Python Document Search

I'm pretty new to Python and only want to extract the city for these clients' addresses:

```
clients = ["Peter, Calle Fantasia 15, Madrid", "Robert, Plaza de Perdas 2,  
Sevilla", "Paul, Calle Polo, Madrid", "Francesco, Plaza de Opo I, Segovia"]
```

Can someone help? Thank you very much in advance!



stackoverflow

```
[i.split(',')[ -1].strip() for i in clients]  
# ['Madrid', 'Sevilla', 'Madrid', 'Segovia']
```

Answer URL

<https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions>

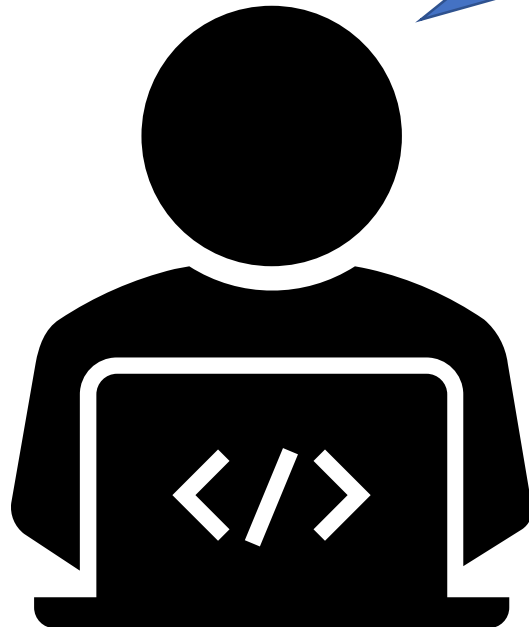
<https://docs.python.org/3/library/stdtypes.html#str.split>

Goal

[Python] Extract certain values from a list



stackoverflow



If the elements of `clients` are always of the format "name, address, city", you can [split](#) it like so:

```
2 # List comprehension, splits each element of client on commas,
# then takes the final element (stripping any whitespace)
clients = [client.split(',')[1].strip() for client in clients]

>>> print(clients)
['Madrid', 'Sevilla', 'Madrid', 'Segovia']
```

[share](#) [improve this answer](#)

[add a comment](#)

4 You can use a list comprehension, and keep the last element in each string starting from the onwards.

For that use [str.strip\(\)](#) setting `.` as a separator, which will split each string whenever comma, slice the resulting lists keeping the last element, and use [string.strip](#) to remove white spaces.

```
clients = ["Peter, Calle Fantasia 15, Madrid", "Robert, Plaza de Perdas 2, Sevilla", "Paul, C. list_of_cities = []
for i in clients:
    index_last_comma = 0
    for j in range(len(i)-1,0,-1):
        if i[j] == ',':
            index_last_comma = j
            break
    city = i[index_last_comma+1:].strip()
    list_of_cities.append(city)

print(list_of_cities)
# output ['Madrid', 'Sevilla', 'Madrid', 'Segovia']
```

[share](#) [improve this answer](#)

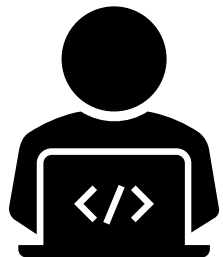
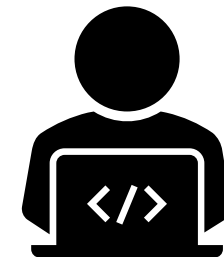
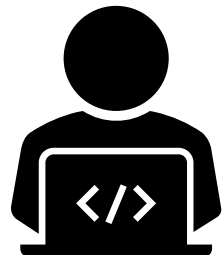
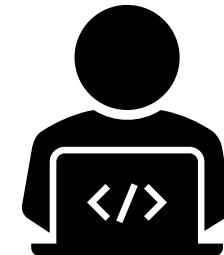
edited Mar 20 at 11:41

answered Mar 20 at 11:37

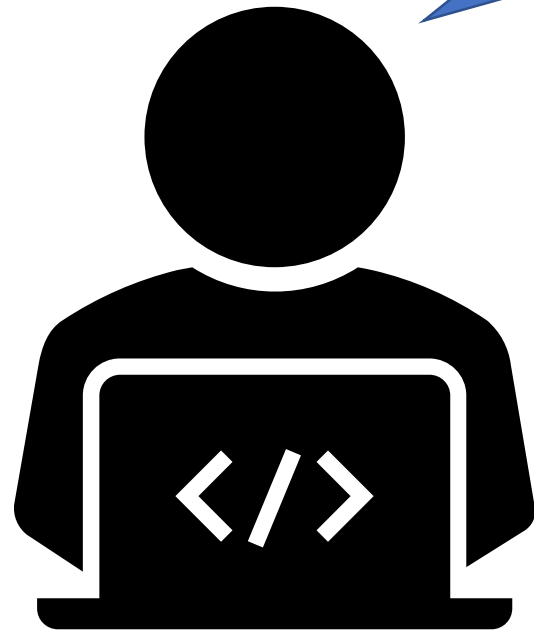
prashant rana
1,169 • 9 • 20

For more details on the methods used above, I'd suggest you give a look at:

- [string — Common string operations](#)
- [List Comprehensions](#)



Goal



[Python] Extract certain values from a list

Look at these URLs

<https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions>

<https://docs.python.org/3/library/stdtypes.html#str.split>

stackoverflow



Contents

- Data mining and Preprocessing
 - Stackoverflow Python Q&A Dataset
 - Python Document
- Actual Works Explained
 - Word/Document Vector Representation
 - Web Interface
- Evaluation
- Future Works

Contents

- **Data mining and Preprocessing**
 - **Stackoverflow Python Q&A Dataset**
 - Python Document
- Actual Works Explained
 - Word/Document Vector Representation
 - Web Interface
- Evaluation
- Future Works

Stackoverflow Python Q&A Dataset

- **47491 Q&As, 87 MB size**
- Some are duplicated Q&A (redirected question)
- Constraints
 - [Python] tagged Q&A questions
 - The page should contain the link to “docs.python.org/3/”
- You can download this dataset at
 - <https://github.com/cushionbadak/PyMaker/tree/master/PyMaker/datas/object3>
- Unrefined data is also available at
 - <https://github.com/cushionbadak/PyMaker/tree/master/PyMaker/datas/object2>
- Stackoverflow’s all user contributions are licensed under [Creative Commons Attribution-Share Alike](#)

Stackoverflow Python Q&A Dataset

```
Python_QNA_URLs_1_100.txt - 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
/questions/55260190/how-to-speed-up-my-keras-cnn-with-pre-trained-inceptionv3
/questions/55260183/make-push-notifications-using-python-bottle-library
/questions/55260179/changing-tmux-window-name-using-python
/questions/55260172/how-to-compare-the-size-of-tf-variable-and-scalar-in-tensorflow
/questions/55260097/python-trace-process-which-is-started-from-python-script
/questions/55260093/how-to-count-number-of-all-child-under-every-p1-id-in-pandas-id-and-parent-id-a
/questions/55260073/python-identify-user-input-positive-or-negative
/questions/55260045/how-can-i-use-a-different-separators-in-join-in-python
/questions/55260041/getting-error-when-running-python-code-on-ubuntu
/questions/55259985/animated-image-with-python
/questions/55259955/pep8-import-guideline-contradicts-principle-of-minimal-variable-span-and-visibility
/questions/55259953/pytest-generate-tests-in-derived-classes
/questions/55259864/only-size-1-arrays-can-be-converted-to-python-scalars
/questions/55259818/how-to-add-django-admin-as-foreign-key-in-django
/questions/55259791/how-to-compare-timestamps-of-2-files-with-different-frequencies
/questions/55259788/how-to-change-y-axis-on-a-pandas-dataframe-plot
/questions/55259774/string-with-no-spaces-needs-to-split-based-on-pattern
/questions/55259773/how-to-get-the-name-of-a-csv-file-causing-an-error-in-dask-read-csv
/questions/55259769/extract-e-mails-from-multiple-pages-in-a-website-and-list-it
/questions/55259755/maximize-number-of-parallel-requests-aihttp
/questions/55259711/is-there-a-command-line-tool-to-automatically-setup-and-upload-my-python-package
/questions/55259682/rendering-folium-objects
/questions/55259666/convert-py-to-exe-using-py2exe-specifically
/questions/55259658/deleting-lines-in-multiple-csv-files-in-a-folder-with-python
/questions/55259637/counting-with-variable-base-for-each-digit
/questions/55259609/python-subprocess-with-double-quote-and-whitespace-for-django-command
/questions/55259601/extract-certain-values-from-a-list
/questions/55259568/google-api-error-credentials-file-doesnt-exist
/questions/55259504/exact-and-case-insensitive-match-for-a-multi-word-token-in-a-string-python
/questions/55259503/ib-api-not-installing-properly-on-mac
/questions/55259500/python-pr0s-how-would-you-clean-this-dictionary
/questions/55259486/hash-computation-vs-bucket-walkthrough
/questions/55259461/how-to-convert-ms-access-query-objects-to-sqlite-views
/questions/55259437/attributeerror-webdriver-object-has-no-attribute-select-by-visible-text-whi
/questions/55259410/can-anybody-explain-the-meaning-of-glob-glob-function-in-python
/questions/55259397/how-to-make-change-of-variables-in-animation-funcanimation-function-in-python
/questions/55259381/how-to-make-the-prediction-in-neural-network
/questions/55259371/pytest-testing-parser-error-unrecognised-arguments
/questions/55259340/matrix-multiplication-in-python-giving-error-how-can-i-overcome-this
/questions/55259313/print-the-local-date-and-time-in-python
/questions/55259269/flask-swagger-api-doc-requires-positional-argument-but-actually-given
/questions/55259255/change-pandas-column-based-on-another-column
/questions/55259221/how-to-pass-a-dopostback-button-to-get-html-in-python
/questions/55259218/thread-pool-is-slow-and-same-speed-as-serial
/questions/55259210/importing-a-file-from-a-subfolder-with-read-csv-how-to-get-it-to-work-with-eng
/questions/55259204/memory-error-when-running-two-spark-jobs-in-succession-no-problem-when-running
/questions/55259183/adding-x-y-line-to-plot-containing-boxplot
/questions/55259175/how-to-open-a-random-video-from-a-playlist-in-py-3-6
```

```
Python_DOC_referPages.extract-certain-values-from-a-list.txt - 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

Natural Text
I'm pretty new to Python and only want to extract the city for these clients' addresses:Can
someone help? Thank you very much in advance!
You can use a list comprehension, and keep the last element in each string starting from the 1st
, onwards. For that use string.split setting , as a separator, which will split each string whenever
there is a comma, slice the resulting lists keeping the last element, and use string.strip to
remove leading white spaces:For more details on the methods used above, I'd suggest you give
a look at:string — Common string operationsList Comprehensions
If the elements of clients are always of the format "name, address, city", you can split it like so:
without using list comprehension

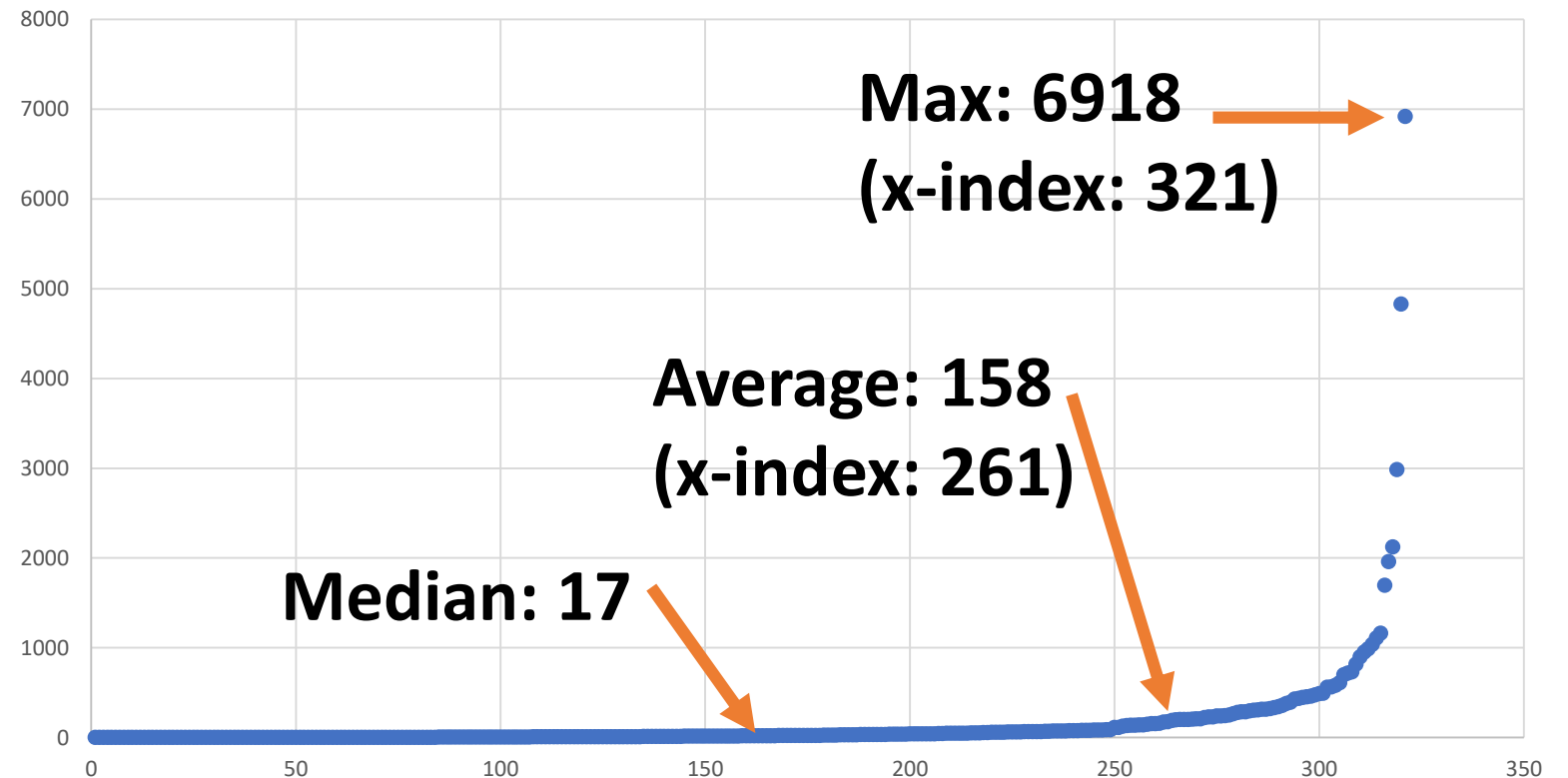
Answer URL
https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions
https://docs.python.org/3/library/stdtypes.html#str.split

city = re.findall(r'\w+', client['address'])
list_of_cities.append(city)

print(list_of_cities)
# output ['Madrid', 'Sevilla', 'Madrid', 'Segovia']
</code> </pre>
</div>
<div class="comment-copy">Is the city always the third part of <code>&lt;name&gt; &lt;street&gt; &lt;city&gt;</code>. Try to <code>split</code> each string.</span>
<span class="comment-copy">Don't forget to give some feedback @lane -, see <a href="https://stackoverflow.com/help/someone-answers">What should I do when someone answers my question?</a></span>
<span class="comment-copy">what is the point of using <code>strip(' ')</code> over <code>strip0</code></span>
<span class="comment-copy">None @prashantrana -></span>
```

Stackoverflow Python Q&A Dataset

- Chart: The Number of times mentioned for each link
 - X – index number of each link
 - Y – # of times
- Small & Biased



Stackoverflow Python Q&A Dataset

- Construct Stackoverflow Python Q&A Dataset
- We need to collect **Python Document itself** too

Contents

- **Data mining and Preprocessing**
 - Stackoverflow Python Q&A Dataset
 - **Python Document**
- Actual Works Explained
 - Word/Document Vector Representation
 - Web Interface
- Evaluation
- Future Works

Collecting Python Document - contents

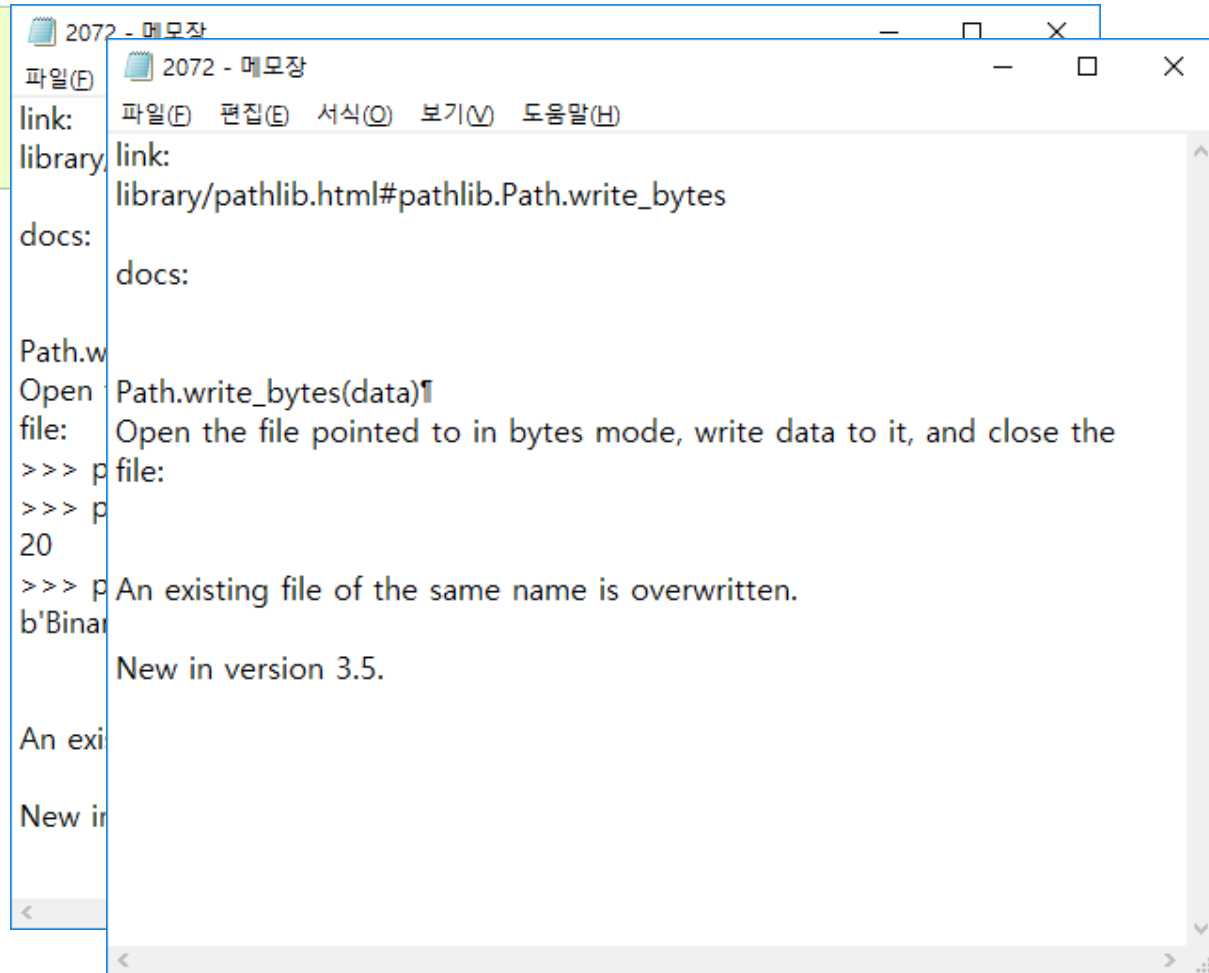
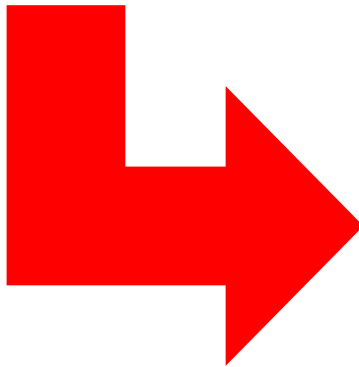
`Path.write_bytes(data)`

Open the file pointed to in bytes mode, write *data* to it, and close the file:

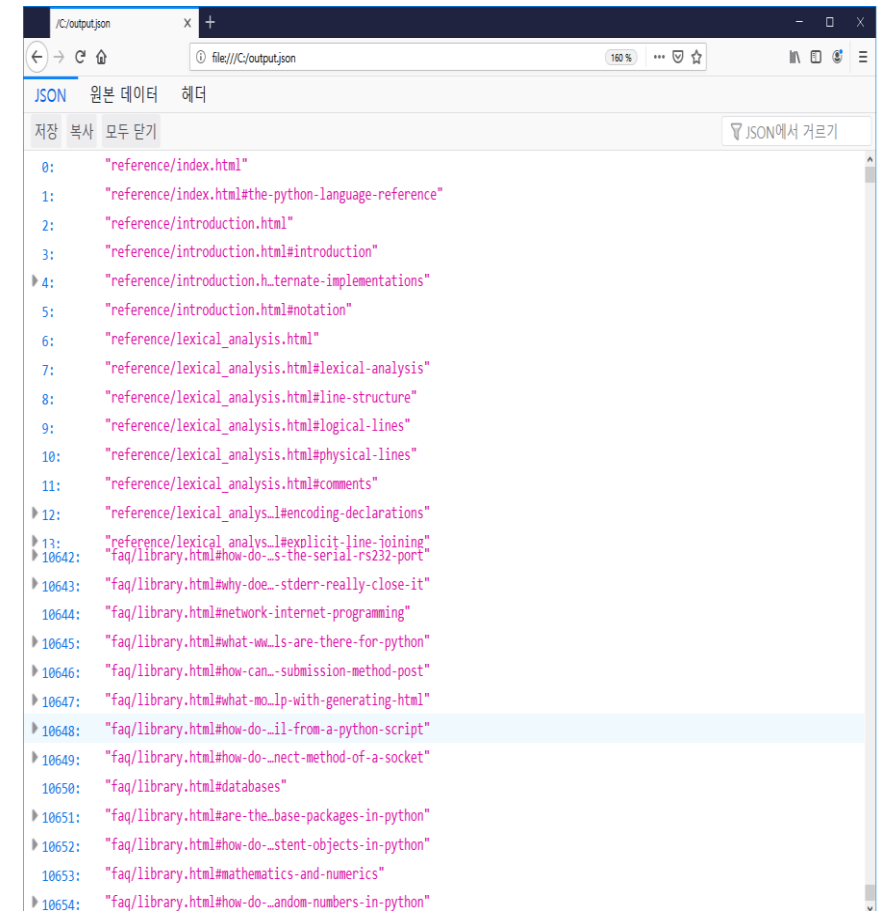
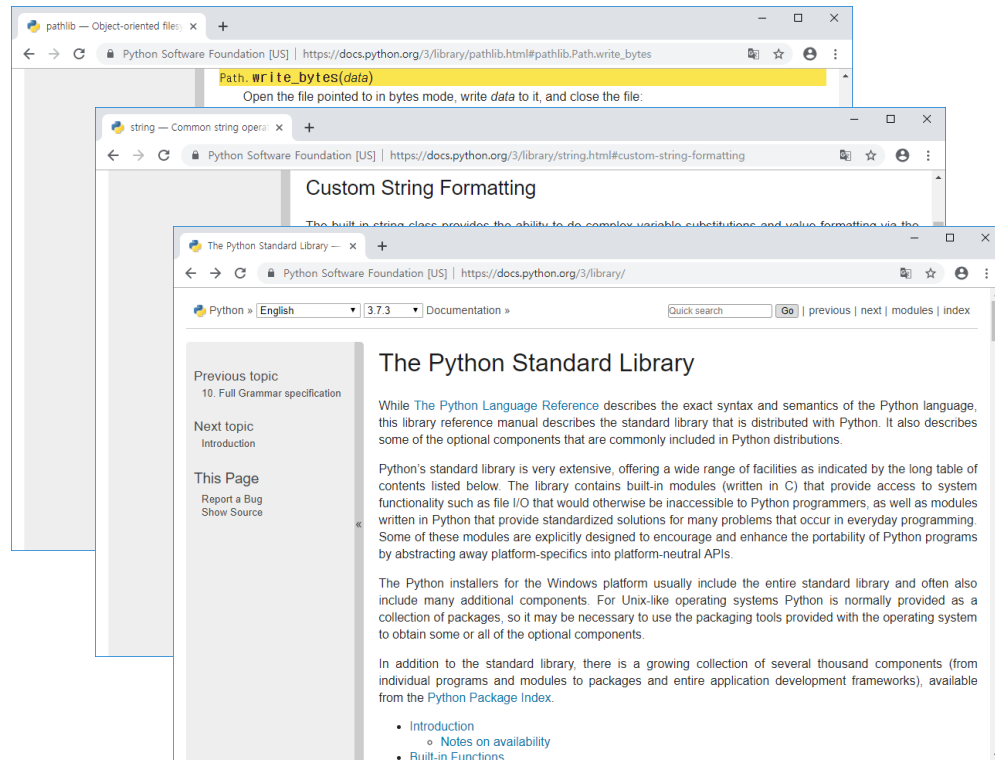
```
>>> p = Path('my_binary_file')
>>> p.write_bytes(b'Binary file contents')
20
>>> p.read_bytes()
b'Binary file contents'
```

An existing file of the same name is overwritten.

New in version 3.5.



Collecting Python Document - URLs



Collecting Python Document

- **10655 Documents, 28 MB size**
- Some are duplicated contents
 - For example, https://docs.python.org/3/reference/lexical_analysis.html#keywords
 - contents are included in https://docs.python.org/3/reference/lexical_analysis.html
- Constraints
 - We only collected URLs and content at <https://docs.python.org/3/library/> and <https://docs.python.org/3/reference/>
- You can download this dataset at
 - <https://github.com/cushionbadak/PyMaker/tree/master/PyMaker/datas/object4>
- Licenses for Python documentation are located at the following links: <https://docs.python.org/3/license.html>

Contents

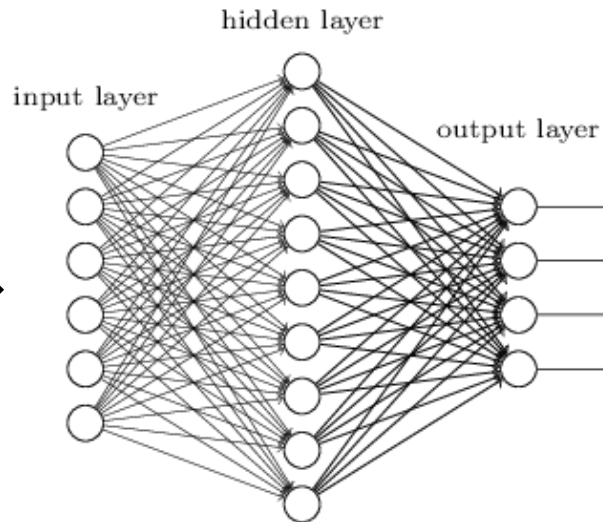
- Data mining and Preprocessing
 - Stackoverflow Python Q&A Dataset
 - Python Document
- Actual Works Explained
 - **Word/Document Vector Representation**
 - Web Interface
- Evaluation
- Future Works

First Attempt

- We believed that the bi-gram sentence classification would solve our problem

Bi-gram representation

hash('Hello-Python') →



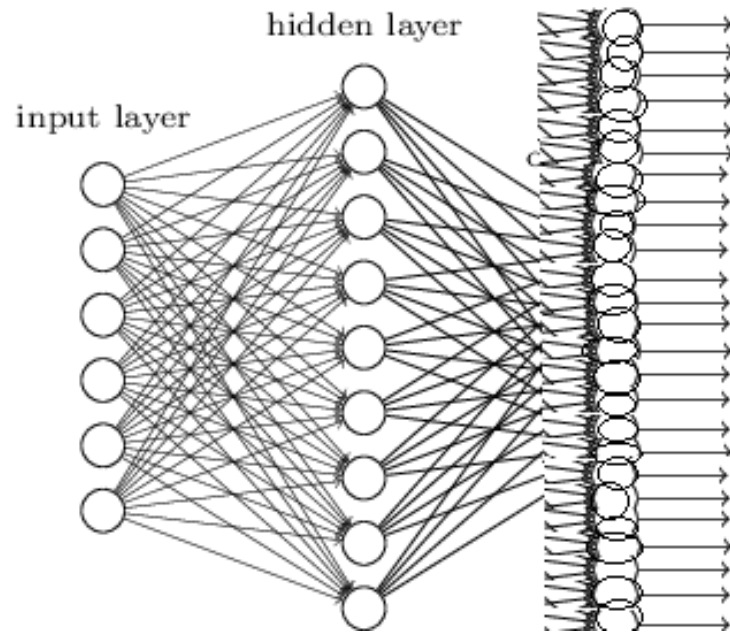
Scores for each URL

→ <1.1, 2.0, -0.7, ..., 0.2>

- Image source: <https://www.extremetech.com/wp-content/uploads/2016/05/tikz35.png>

First Attempt - Problem

1. The size of class



There are **10655** URLs in
<https://docs.python.org/3>

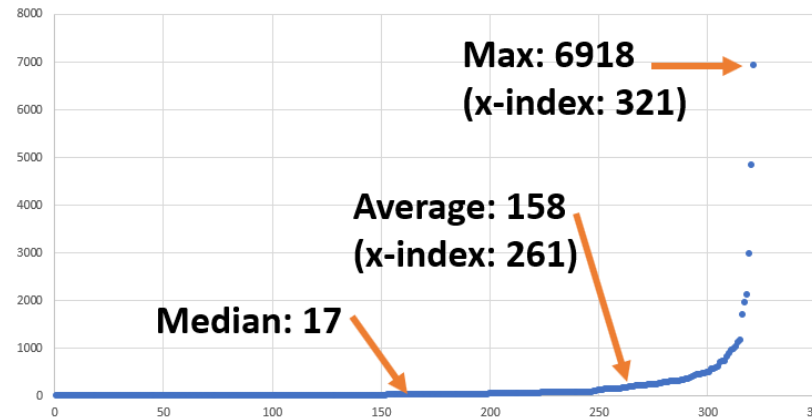
So we reduced to **321** html file URLs

First Attempt - Problem

2. The size of training data

Stackoverflow Python Q&A Dataset

- Chart: The Number of times mentioned for each link.
 - X – index number of each link
 - Y – # of times
- Small & Biased



First Attempt

```
result_so_zerobase_learned_full_r.log
1 ITEATION 1 => CORRECT: 0 / 0 FILE: break-out-of-a-python-for-loop-using-boolean-flag-logic-and-indentation-error
2 CANDIDATES:
3 library/tkinter.scrolledtext.html
4 library/quopri.html
5 library/array.html
6 reference/simple_stmts.html
7 library/__future__.html
8 ANSWERS:
9
10 ITEATION 2 => CORRECT: 0 / 1 FILE: pygame-load-images-that-change-filename
11 CANDIDATES:
12 library/tkinter.scrolledtext.html
13 library/quopri.html
14 reference/simple_stmts.html
15 library/array.html
16 library/spwd.html
17 ANSWERS:
18 library/os.html
19
20 ITEATION 3 => CORRECT: 0 / 1 FILE: write-filenames-of-different-extention-into-different-text-files
21 CANDIDATES:
22 library/tkinter.scrolledtext.html
23 library/quopri.html
24 reference/simple_stmts.html
25 library/array.html
26 library/email.errors.html
27 ANSWERS:
28 library/functions.html
29
30 ITEATION 4 => CORRECT: 0 / 1 FILE: getting-typeerror-when-attempting-to-open-chrome-using-webbrowser-python
31 CANDIDATES:
32 library/tkinter.scrolledtext.html
33 library/quopri.html
34 reference/simple_stmts.html
35 library/array.html
36 library/email.errors.html
37 ANSWERS:
38 library/webbrowser.html
39
40 ITEATION 5 => CORRECT: 0 / 1 FILE: how-do-i-remove-whitespace-to-balance
41 CANDIDATES:
```

```
result_so_zerobase_learned_full_r2.log
5034 library/itertools.html
5035 library/python.html
5036 library/math.html
5037 ANSWERS:
5038 library/asyncio-eventloop.html
5039 library/asyncio-task.html
5040 library/concurrent.futures.html
5041
5042 ITEATION 498 => CORRECT: 0 / 2 FILE: unexpected-long-decimal-after-converting-to-float
5043 CANDIDATES:
5044 library/faulthandler.html
5045 library/io.html
5046 library/python.html
5047 library/itertools.html
5048 library/math.html
5049 ANSWERS:
5050 library/functions.html
5051 library/decimal.html
5052
5053 ITEATION 499 => CORRECT: 0 / 1 FILE: utf-8-codec-cant-decode-byte-0xb5-in-position-0-invalid-start-byte
5054 CANDIDATES:
5055 library/faulthandler.html
5056 library/itertools.html
5057 library/io.html
5058 library/python.html
5059 library/math.html
5060 ANSWERS:
5061 library/codecs.html
5062
5063 ITEATION 500 => CORRECT: 0 / 1 FILE: python-unittest-assert-called-with-false-despite-identical-calls
5064 CANDIDATES:
5065 library/faulthandler.html
5066 library/io.html
5067 library/python.html
5068 library/itertools.html
5069 library/xml.etree.elementtree.html
5070 ANSWERS:
5071 library/unittest.mock.html
5072
5073 TOTAL CORRECT: 35 / 572
5074
```

Second Attempt

- Since our computing resource is limited, we decided to use pre-trained word2vec model
- We use GoogleNews-vectors-negative300.bin (**3.6 GB, 300-dimensional, 3 million words and phrases**, trained with 100 billion Google News dataset)
 - <https://code.google.com/archive/p/word2vec/>
- We use gensim library to load model
 - <https://radimrehurek.com/gensim/>

Second Attempt

< Classification Algorithm >

1. Convert every python document into vectors
 - **The vector is obtained by adding all the words in the document as vectors**

Second Attempt

< Classification Algorithm >

1. Convert every python document into vectors
 - **The vector is obtained by adding all the words in the document as vectors**
2. Convert query string into vector
 - Using the same method as 1

Second Attempt

< Classification Algorithm >

1. Convert every python document into vectors
 - **The vector is obtained by adding all the words in the document as vectors**
2. Convert query string into vector
 - Using the same method as 1
3. Find the most similar python document using **cosine similarity**

Contents

- Data mining and Preprocessing
 - Stackoverflow Python Q&A Dataset
 - Python Document
- Actual Works Explained
 - Word/Document Vector Representation
 - **Web Interface**
- Evaluation
- Future Works

Web Interface

- You can run the current version of our document search program at <http://13.125.156.35:8000/Query/> (Open temporarily)

PyMaker

ex) returns the time in seconds

urns the time in seconds

Type: #1 ☒ #2 ☐

Query

Web Interface

- Result Page

Python Reference to "returns the time in seconds"

- <https://docs.python.org/3/library/time.html>
- <https://docs.python.org/3/library/datetime.html>
- <https://docs.python.org/3/library/calendar.html>
- <https://docs.python.org/3/library/sched.html>
- <https://docs.python.org/3/library/asyncio-queue.html>

[New Query](#)

Contents

- Data mining and Preprocessing
 - Stackoverflow Python Q&A Dataset
 - Python Document
- Actual Works Explained
 - Word/Document Vector Representation
 - Web Interface
- **Evaluation**
- Future Works

Evaluation

- There are no well-known benchmarks for python
- So we measured our algorithm's classification accuracy with Stackoverflow Python Q&A Dataset
- For now, our model successfully classifies **5781 of 41568 (13.9%)**
 - The model gives 5-candidates for each given Question-title sentence
 - 5 Random selection will show only 1.6% of accuracy
- We hope to evaluate the accuracy of other web search engines on this Stackoverflow benchmark
 - It is hard to gather the result pages of search engines

Contents

- Data mining and Preprocessing
 - Stackoverflow Python Q&A Dataset
 - Python Document
- Actual Works Explained
 - Word/Document Vector Representation
 - Web Interface
- Evaluation
- **Future Works**

Future Works

- We will continue to explore other methods, models for better document-to-vector representation
- If possible, we will evaluate our Stackoverflow Q&A benchmark on other search engines
- We will increase the usability of our Python document search engine

Summary

Contents

- Data mining and Preprocessing
 - Stackoverflow Python Q&A Dataset
 - Python Document
- Actual Works Explained
 - Word/Document Vector Representation
 - Web Interface
- Evaluation
- Future Works

Summary

- Creativity
 - It's Python Document Search Engine!
 - We made our own benchmark to evaluate this problem
- Technical Completeness
 - Document classification: Base 1.6% → Our approach 13.9%
- Contribution
 - We publicly open our source codes and data at github
 - We provide web services to better understand the project

Q&A

감사합니다