자연어처리 프로젝트 3차발표

2조

2013210043 권민규

2014210035 전수혁

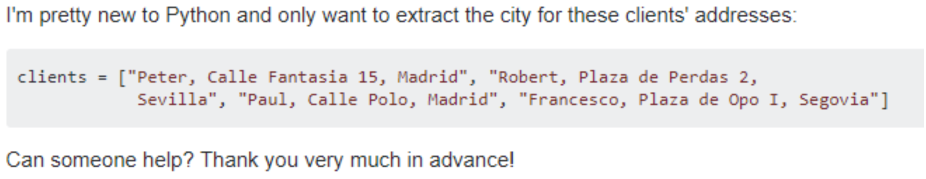
2014210064 변지석

2019/June/12

Python Document Search

파이썬 문서 검색기

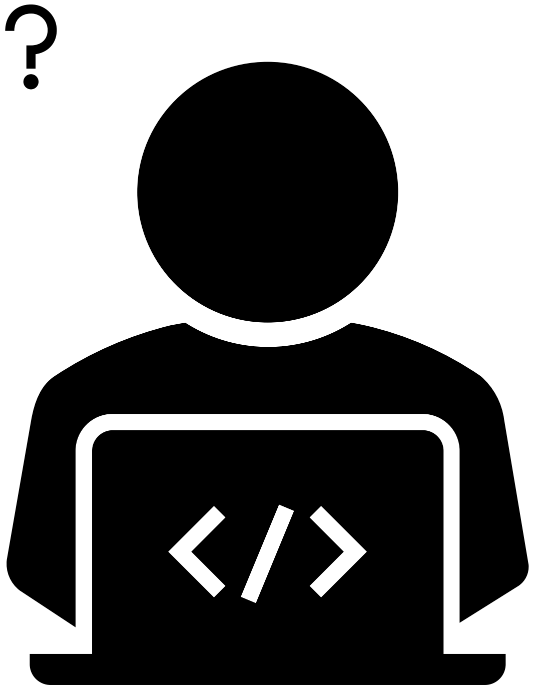
Why Python Document Search



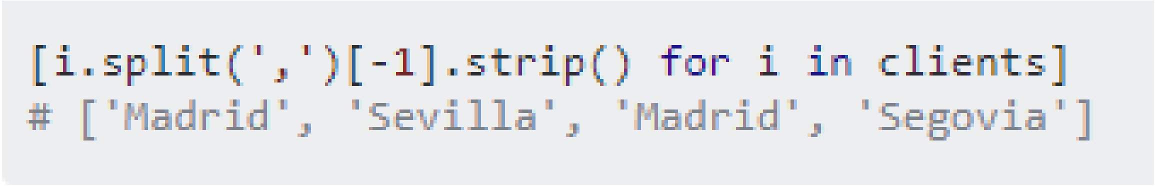
I'm pretty new to Python and only want to extract the city for these clients' addresses:

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

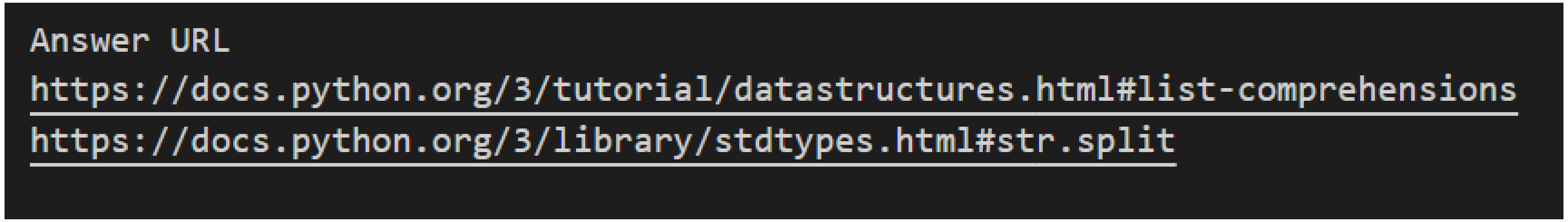
Can someone help? Thank you very much in advance!



stackoverflow



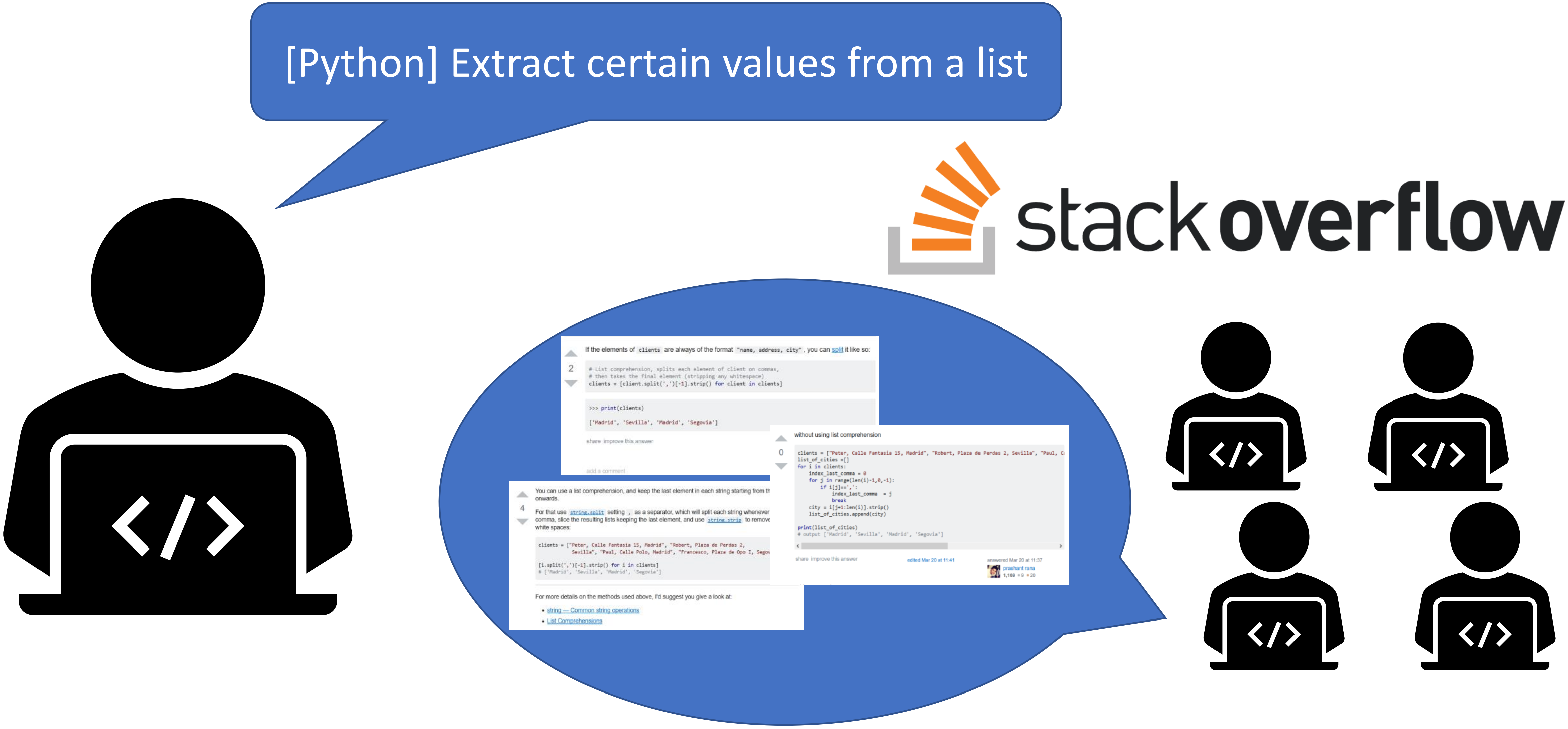
```
[i.split(',')[-1].strip() for i in clients]
# ['Madrid', 'Sevilla', 'Madrid', 'Segovia']
```



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

Question Source: https://stackoverflow.com/questions/55259601/extract-certain-values-from-a-list

Goal



[Python] Extract certain values from a list

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

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

4

For that use `string.split` 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, Calle Polo, Madrid", "Francesco, Plaza de Opo I, Segov

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

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

- string — Common string operations
- List Comprehensions

without using list comprehension

0

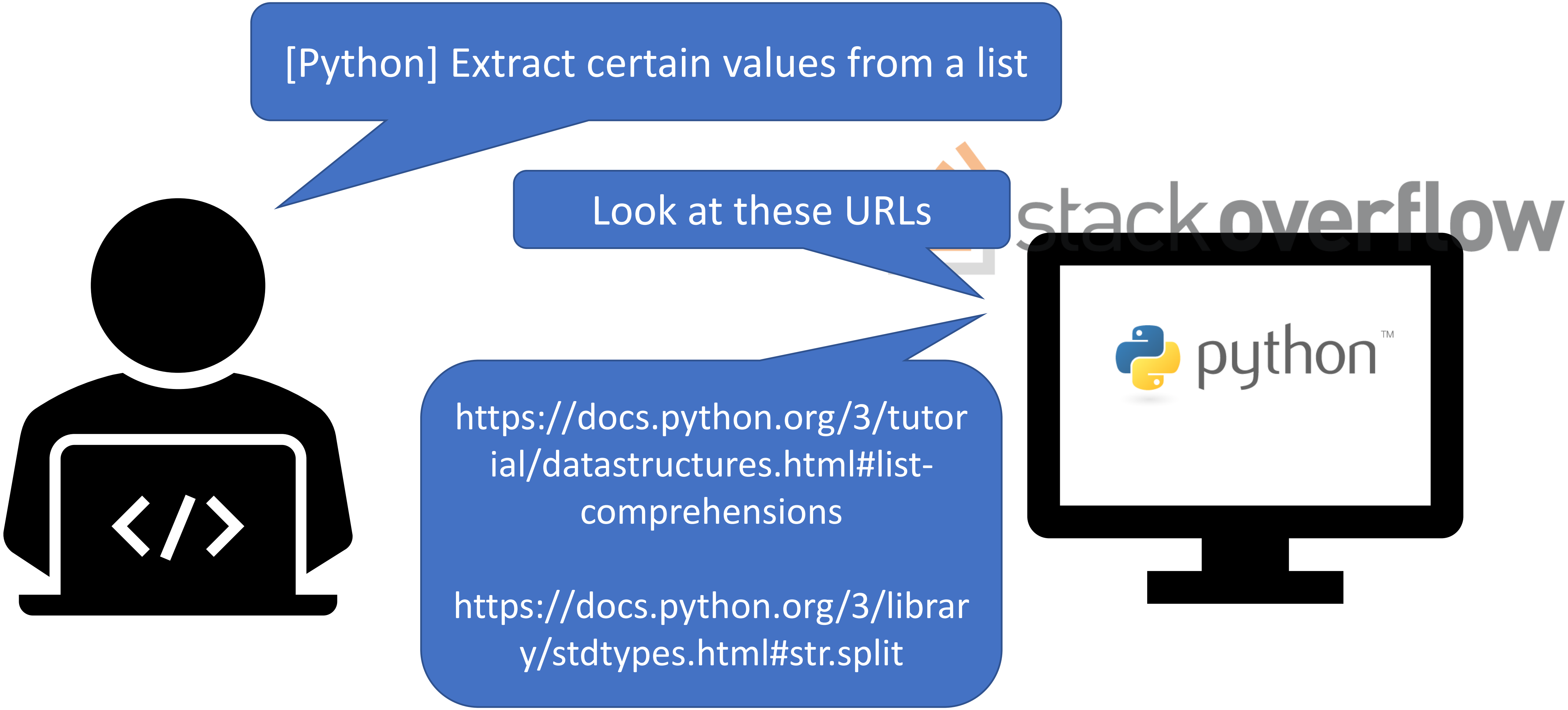
```
clients = ["Peter, Calle Fantasia 15, Madrid", "Robert, Plaza de Perdas 2, Sevilla", "Paul, C
list_of_cities =[]
for i in clients:
    index_last_comma = 0
    for j in range(len(i)-1,0,-1):
        if i[j]==',':
            index_last_comma  = j
            break
    city = i[j+1:len(i)].strip()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

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

Contents

- Search Overview
- Problem & TF-IDF
- Evaluation
- Conclusion

Contents

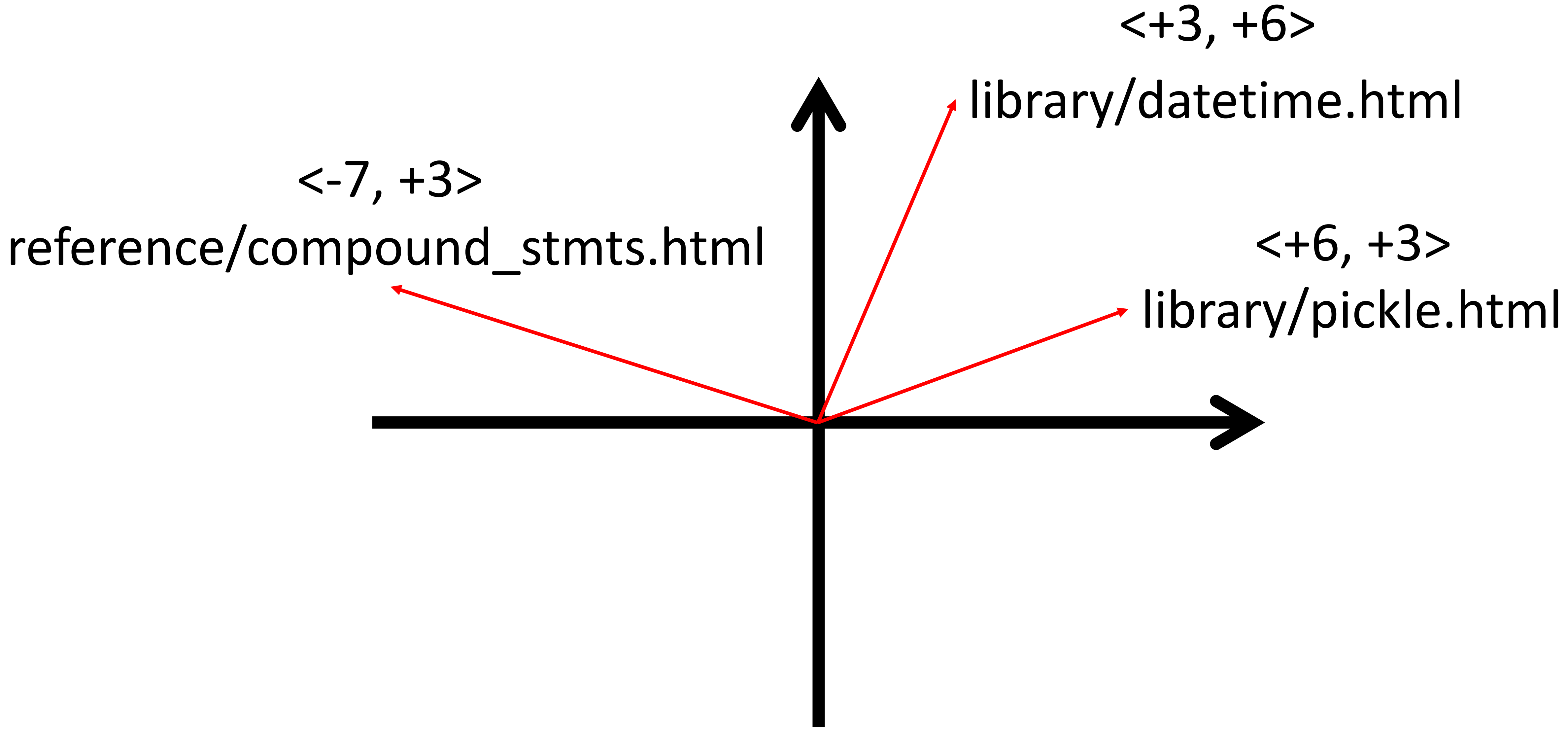
- **Search Overview**
- Problem & TF-IDF
- Evaluation
- Conclusion

Search Overview

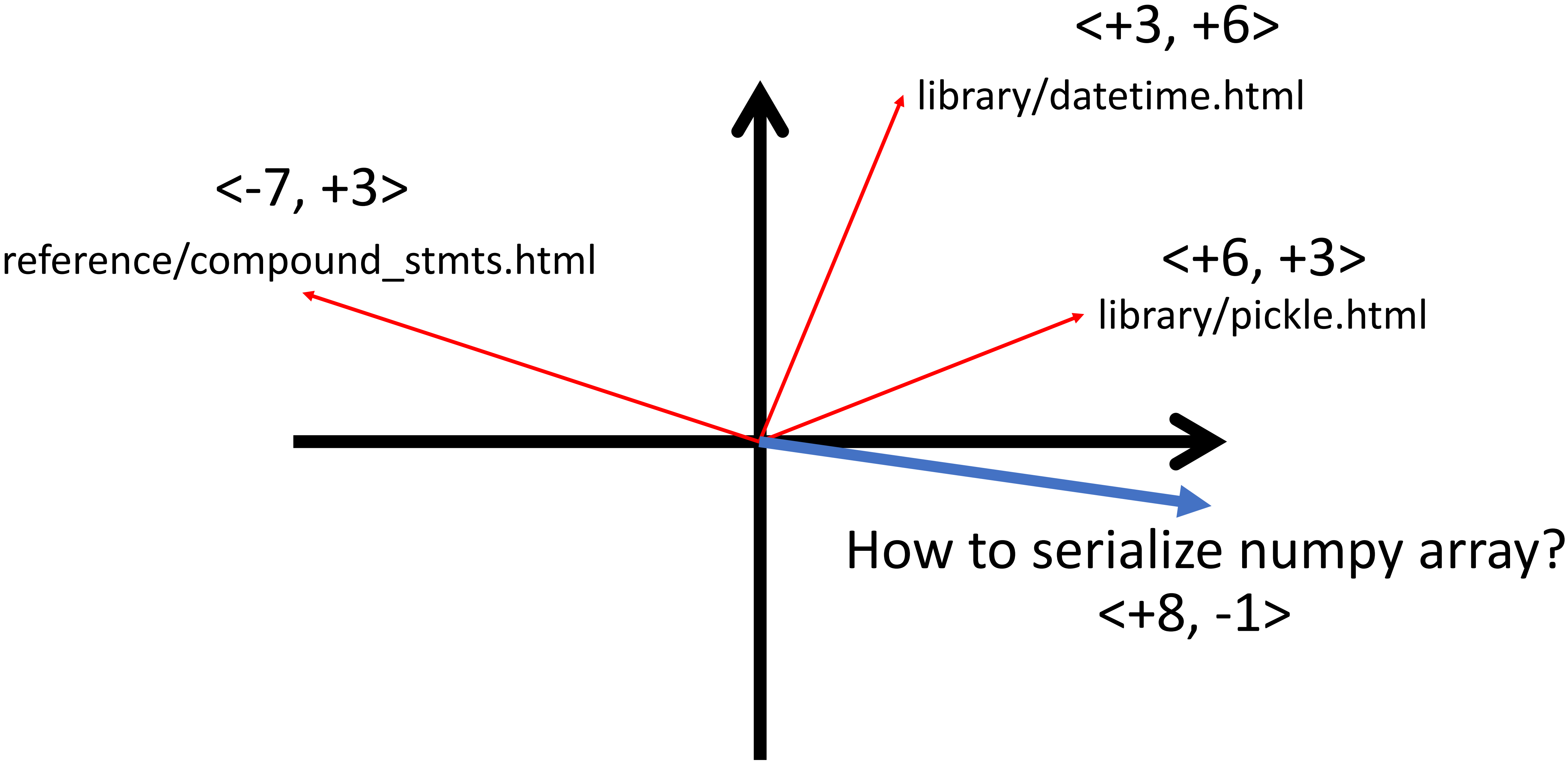
- 분류목표
  - 파이썬 공식 문서 321개

- 각 문서를 벡터로 표현하고 질문 문장과 가장 유사한 벡터를 찾아주자!

Search Overview



Search Overview



<+3, +6>

library/datetime.html

<-7, +3>

reference/compound_stmts.html

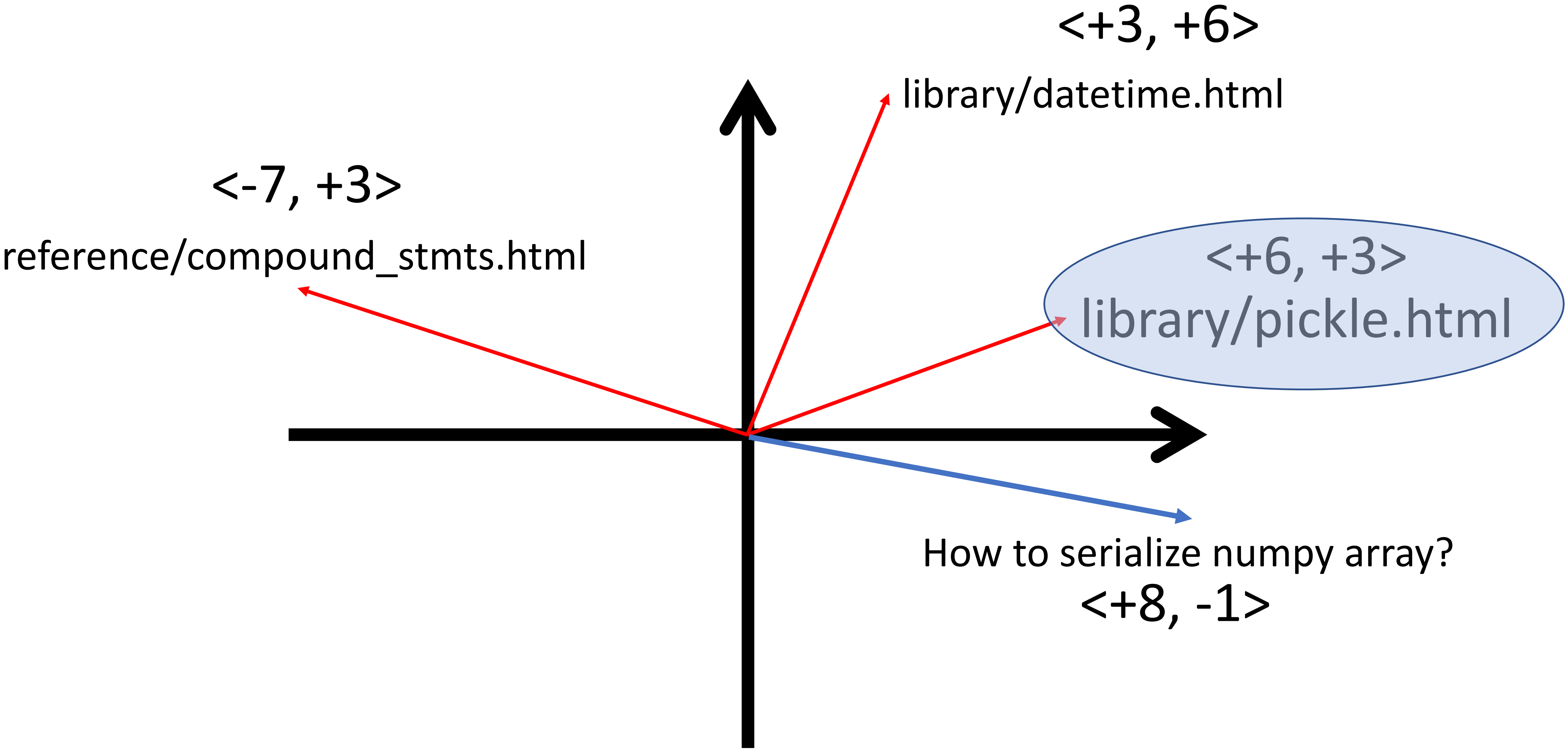
<+6, +3>

library/pickle.html

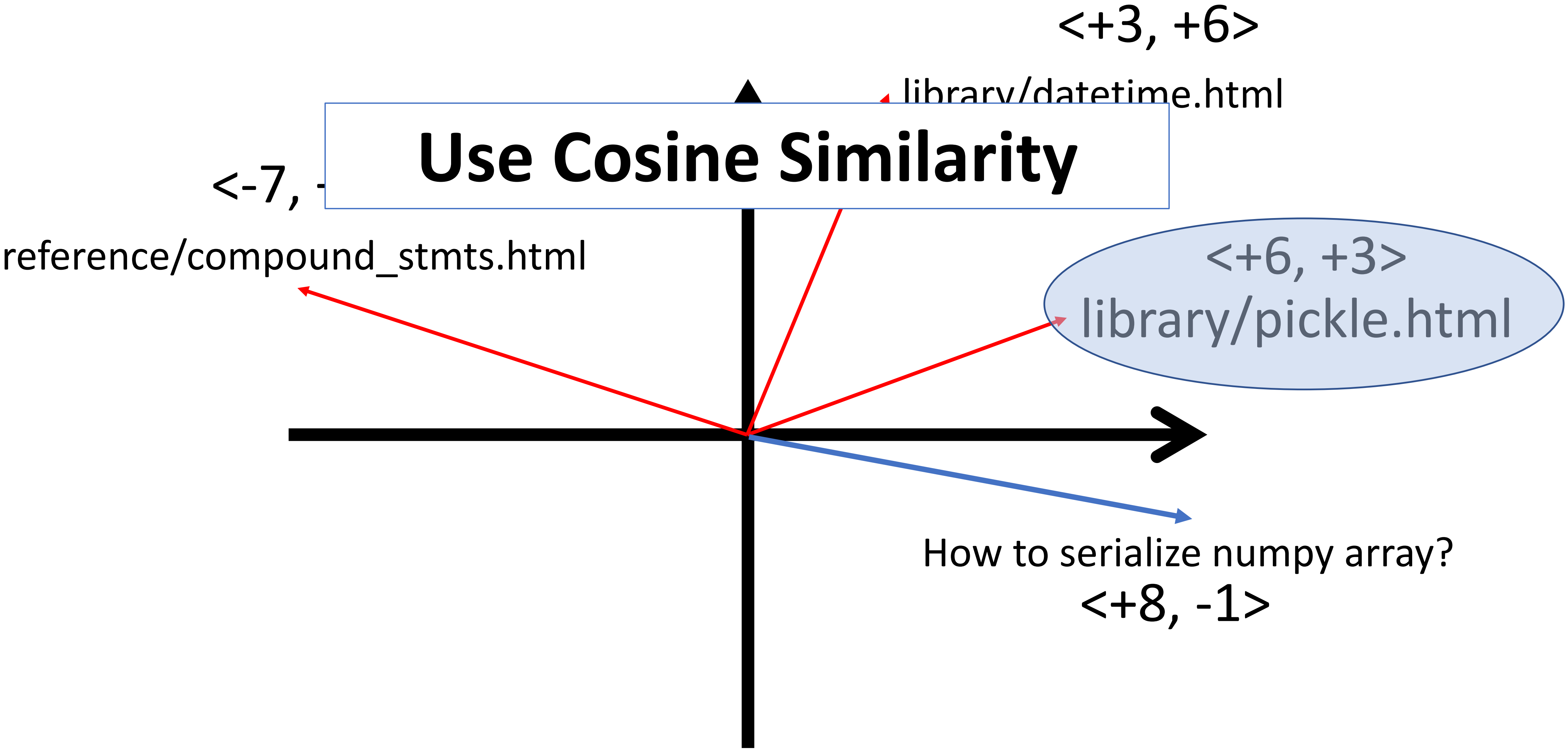
How to serialize numpy array?

<+8, -1>

Search Overview



Search Overview



Use Cosine Similarity

Contents

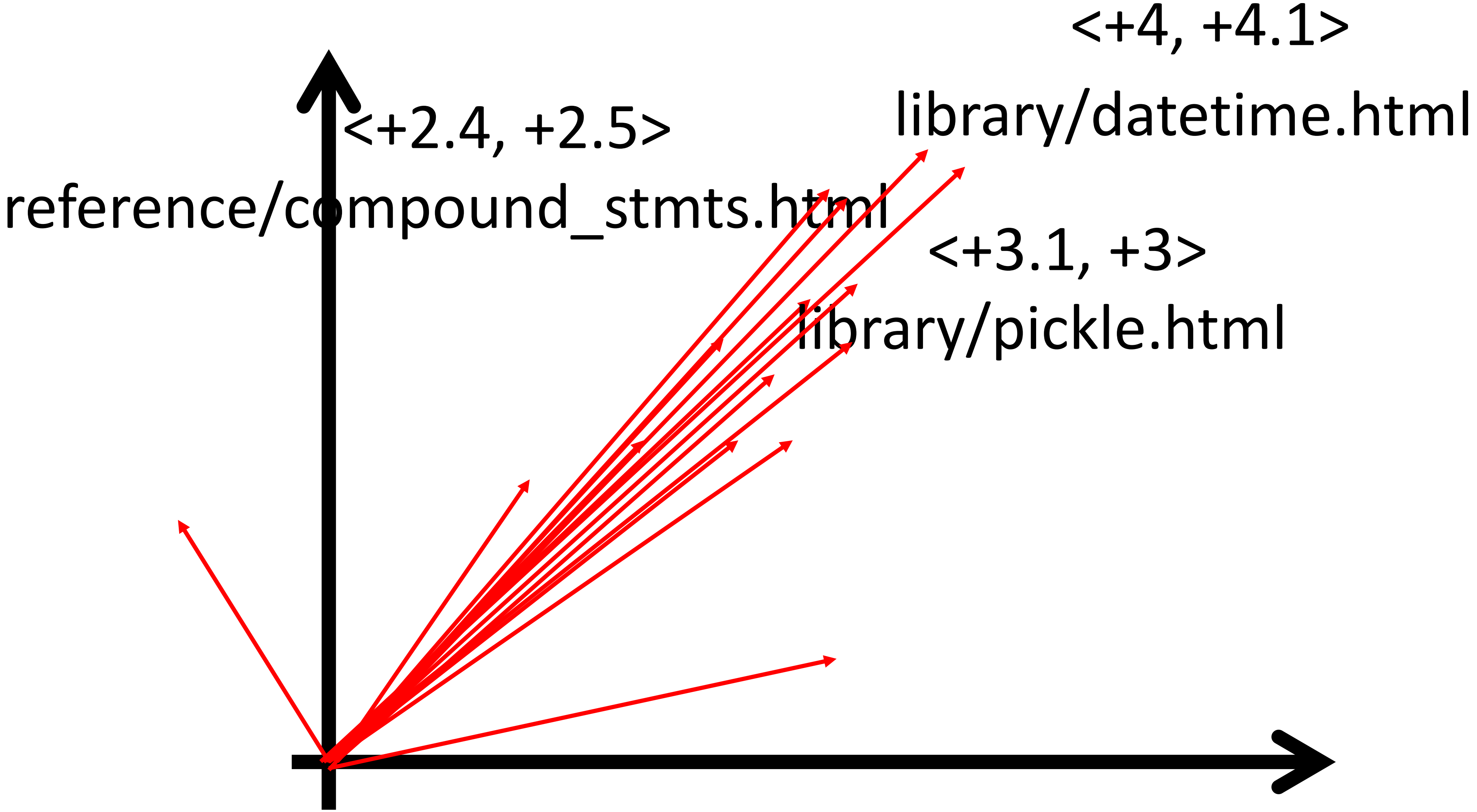
- Search Overview
- Problem & TF-IDF
- Evaluation
- Conclusion

Previous Problem

- Document Vector를 만들 때, 미리 학습된 GoogleNews 단어 벡터 표현을 사용했다.

- 하지만 파이썬 공식 문서에 쓰인 단어들의 벡터 표현 합은 다들 비슷했다.

Previous Problem



TF-IDF

- TF-IDF = log (TF / DF)
  - TF: Term Frequency
  - 그 단어가 한 문서에서 몇 번 나왔는가
  - DF: Document Frequency
  - 그 단어를 포함한 문서가 얼마나 많은가

Document Vector

- 기존: GoogleNews-vectors-negative300.bin.gz 를 활용, 문서에서 등장한 각 단어의 word embedding 값을 그대로 더했다.

- 개선: Document Vector / Query Vector 만들때 TF-IDF 활용

1. Document Vector 만들때 TF-IDF 미적용
2. Document Vector 만들때 TF-IDF 적용
3. Document Vector 만들때 TF-IDF 적용 + DF값 한번 더 적용

a. Query Vector 만들때 DF 미적용
b. Query Vector 만들때 DF 적용

http://34.239.149.222:8000/Query/

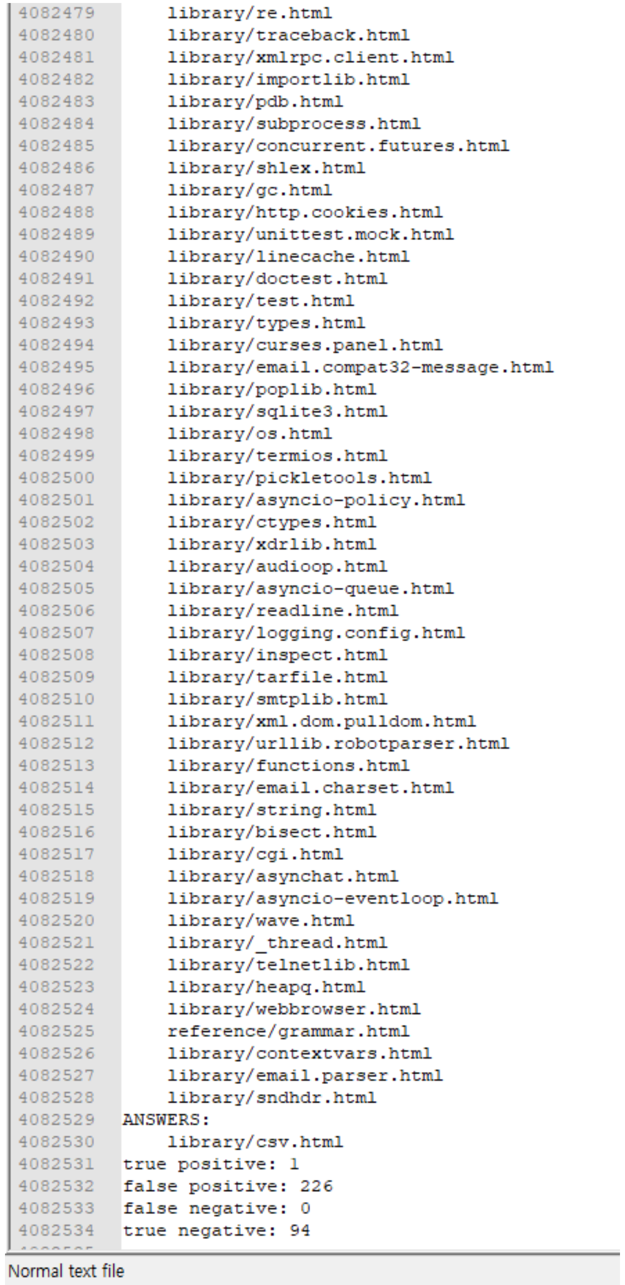
*Open until 6/14

2차

	1+a	1+b	2+a	2+b	3+a	3+b
True Positive	23731	21054	19308	12128	12307	8432
False Positive	3778296	2868162	1761056	1358517	537338	385768
False Negative	17835	20512	22258	29438	29259	33134
True Negative	7521068	8431202	9538308	9940847	10762026	10913596
Precision	0.00624	0.00729	0.01084	0.00885	0.02239	0.02139
Recall	0.57092	0.50652	0.46451	0.29178	0.29608	0.20286
Accuracy	0.66527	0.74529	0.84275	0.87762	0.95004	0.96306

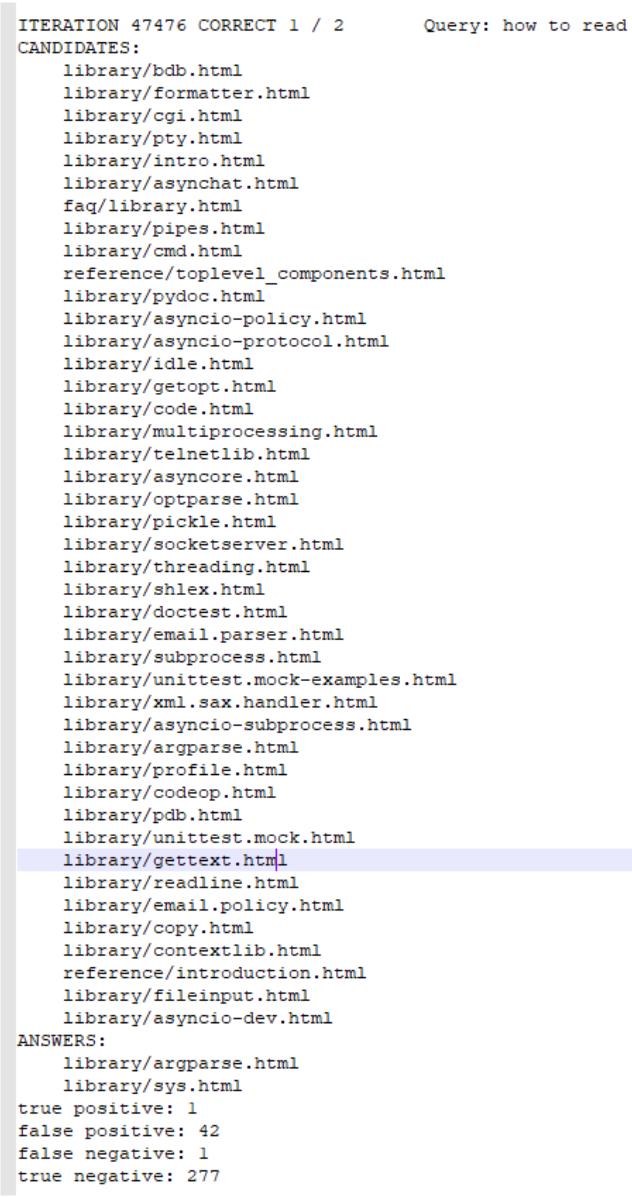
Precision – TP / (TP + FP) , Recall - TP / (TP + FN), Accuracy - (TP + FN) / (TP + FP + TN + FN)

1+a



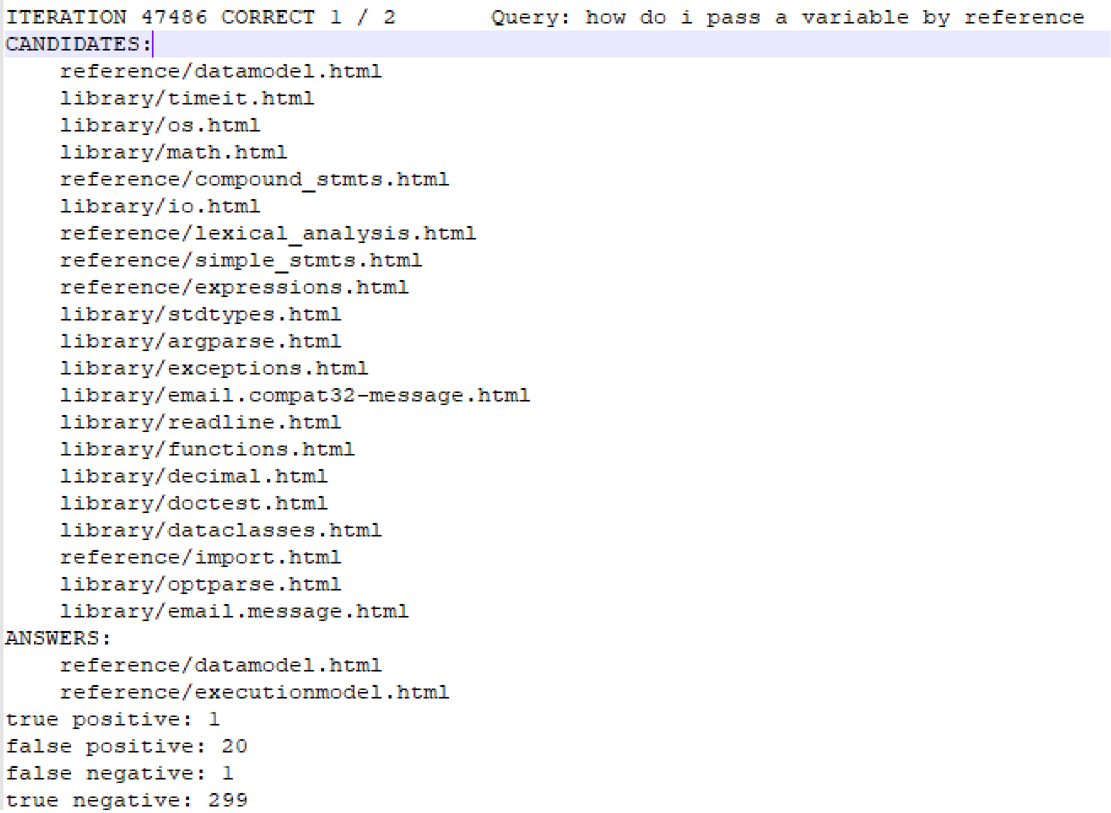
```
    library/re.html
    library/traceback.html
    library/xmlrpc.client.html
    library/importlib.html
    library/pdb.html
    library/subprocess.html
    library/concurrent.futures.html
    library/shlex.html
    library/gc.html
    library/http.cookies.html
    library/unittest.mock.html
    library/linecache.html
    library/doctest.html
    library/test.html
    library/types.html
    library/curses.panel.html
    library/email.compat32-message.html
    library/poplib.html
    library/sqlite3.html
    library/os.html
    library/termios.html
    library/pickletools.html
    library/asyncio-policy.html
    library/ctypes.html
    library/xdrlib.html
    library/audioop.html
    library/asyncio-queue.html
    library/readline.html
    library/logging.config.html
    library/inspect.html
    library/tarfile.html
    library/smtplib.html
    library/xml.dom.pulldom.html
    library/urllib.robotparser.html
    library/functions.html
    library/email.charset.html
    library/string.html
    library/bisect.html
    library/cgi.html
    library/asynchat.html
    library/asyncio-eventloop.html
    library/wave.html
    library/_thread.html
    library/telnetlib.html
    library/heapq.html
    library/webbrowser.html
    reference/grammar.html
    library/contextvars.html
    library/email.parser.html
    library/sndhdr.html
ANSWERS:
    library/csv.html
true positive: 1
false positive: 226
false negative: 0
true negative: 94
```

2+b



```
ITERATION 47476 CORRECT 1 / 2       Query: how to read process command line arguments
CANDIDATES:
    library/bdb.html
    library/formatter.html
    library/cgi.html
    library/pty.html
    library/intro.html
    library/asynchat.html
    faq/library.html
    library/pipes.html
    library/cmd.html
    reference/toplevel_components.html
    library/pydoc.html
    library/asyncio-policy.html
    library/asyncio-protocol.html
    library/idle.html
    library/getopt.html
    library/code.html
    library/multiprocessing.html
    library/telnetlib.html
    library/asyncore.html
    library/optparse.html
    library/pickle.html
    library/socketserver.html
    library/threading.html
    library/shlex.html
    library/doctest.html
    library/email.parser.html
    library/subprocess.html
    library/unittest.mock-examples.html
    library/xml.sax.handler.html
    library/asyncio-subprocess.html
    library/argparse.html
    library/profile.html
    library/codeop.html
    library/pdb.html
    library/unittest.mock.html
    library/gettext.html
    library/readline.html
    library/email.policy.html
    library/copy.html
    library/contextlib.html
    reference/introduction.html
    library/fileinput.html
    library/asyncio-dev.html
ANSWERS:
    library/argparse.html
    library/sys.html
true positive: 1
false positive: 42
false negative: 1
true negative: 277
```

3+b



```
ITERATION 47486 CORRECT 1 / 2       Query: how do i pass a variable by reference
CANDIDATES:
    reference/datamodel.html
    library/timeit.html
    library/os.html
    library/math.html
    reference/compound_stmts.html
    library/io.html
    reference/lexical_analysis.html
    reference/simple_stmts.html
    reference/expressions.html
    library/stdtypes.html
    library/argparse.html
    library/exceptions.html
    library/email.compat32-message.html
    library/readline.html
    library/functions.html
    library/decimal.html
    library/doctest.html
    library/dataclasses.html
    reference/import.html
    library/optparse.html
    library/email.message.html
ANSWERS:
    reference/datamodel.html
    reference/executionmodel.html
true positive: 1
false positive: 20
false negative: 1
true negative: 299
```

Q&A

감사합니다