Introduction to Web Science

Assignment 5

Prof. Dr. Steffen Staab
staab@uni-koblenz.de

René Pickhardt
rpickhardt@uni-koblenz.de

Korok Sengupta
koroksengupta@uni-koblenz.de

Institute of Web Science and Technologies
Department of Computer Science
University of Koblenz-Landau

Submission until: November 30, 2016, 10:00 a.m.
Tutorial on: December 2, 2016, 12:00 p.m.

Please look at the lessons 1) **Dynamic Web Content** & 2) **How big is the Web?**

For all the assignment questions that require you to write code, make sure to include the code in the answer sheet, along with a separate python file. Where screen shots are required, please add them in the answers directly and not as separate files.

Team Name: hotel

Andrea Mildes - mildes@uni-koblenz.de

Sebastian Blei - sblei@uni-koblenz.de

Johannes Kirchner - jkirchner@uni-koblenz.de

Abdul Afghan - abdul.afghan@outlook.de

1 Creative use of the Hyptertext Transfer Protocol (10 Points)

HTTP is a request response protocol. In that spirit a client opens a TCP socket to the server, makes a request, and the server replies with a response. The server will just listen on its open socket but cannot initiate a conversation with the client on its own.

However you might have seen some interactive websites which notify you as soon as something happens on the server. An example would be Twitter. Without the need for you to refresh the page (and thus triggering a new HTTP request) they let you know that there are new tweets available for you. In this exercise we want you to make sense of that behaviour and try to reproduce it by creative use of the HTTP protocol.

Have a look at `server.py`[1] and `webclient.html` (which we provide). Extend both files in a way that after `webclient.html` is servered to the user the person controlling the server has the chance to make some input at its commandline. This input should then be send to the client and displayed automatically in the browser without requiring a reload. For that the user should not have to interact with the webpage any further.

1.1 webclient.html

```
1: <html>
2: <head>
3:         <title>Abusing the HTTP protocol - Example</title>
4: </head>
5: <body>
6:         <h1>Display data from the Server</h1>
7:         The following line changes on the servers command line
8:         input: <br>
9:         <span id="response" style="color:red">
10:                This will be replaced by messages from the server
11:        </span>
12: </body>
13: </html>
```

1.2 Hints:

- This exercise is more like a riddle. Try to focuse on how TCP sockets and HTTP work and how you could make use of that to achieve the expected behaviour. Once you have an idea the programming should be straight-forward.
- The Javascript code that you need for this exercise was almost completely shown in one of the videos and is available on Wikiversity.

[1] you could store the code from http://blog.wachowicz.eu/?p=256 in a file called `server.py`

- In that sense we only ask for a ”proof of concept” nothing that would be stable out in the wilde.
  - In particular, don't worry about making the server uses multithreading. It is ok to be blocking for the sake of this exercise.
- Without use of any additional libraries or AJAX framework we have been able to solve this with 19 lines of Javascript and 11 lines of Python code (we provide this information just as a way for you to estimate the complexity of the problem, don't worry about how many lines your solution uses).

Answer:
webclient

```
<!--assignment 5 task 1
# Andrea Mildes - mildes@uni-koblenz.de
# Sebastian Blei - sblei@uni-koblenz.de
# Johannes Kirchner - jkirchner@uni-koblenz.de
#Abdul Afghan - abdul.afghan@outlook.de -->
<html>
<head>
        <title>Abusing the HTTP protocol - Example</title>
        <script>
        var ajax = new XMLHttpRequest();
        ajax.open('GET', 'http://localhost:8080/webclient.html?long=1', true);
        ajax.onreadystatechange = function () {
            var response_text = ajax.responseText;
            var bodyHtml = /<body.*?>([\s\S]*)<\/body>/.exec(response_text)[1];
            document.getElementById("response").innerHTML = bodyHtml;
        };
        ajax.send(null);
    </script>
</head>
<body>
        <h1>Display data from the Server</h1>
        The following line changes on the servers command line
        input: <br>
        <span id="response" style="color:red">
                This will be replaced by messages from the server
        </span>
</body>
</html>
```

server

```
#!/usr/bin/python
# assignment 5 task 1
# Andrea Mildes - mildes@uni-koblenz.de
# Sebastian Blei - sblei@uni-koblenz.de
# Johannes Kirchner - jkirchner@uni-koblenz.de
```

```
# Abdul Afghan - abdul.afghan@outlook.de

import socket  # Networking support
import signal  # Signal support (server shutdown on signal receive)
import time    # Current time

class Server:
 """ Class describing a simple HTTP server objects."""

 def __init__(self, port = 80):
     """ Constructor """
     self.host = ''   # <-- works on all avaivable network interfaces
     self.port = port
     self.www_dir = 'www' # Directory where webpage files are stored

 def activate_server(self):
     """ Attempts to aquire the socket and launch the server """
     self.socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
     self.socket.setsockopt(socket.SOL_SOCKET, socket.SO_REUSEADDR, 1)
     try: # user provided in the __init__() port may be unavaivable
         print("Launching HTTP server on ", self.host, ":",self.port)
         self.socket.bind((self.host, self.port))

     except Exception as e:
         print ("Warning: Could not aquite port:",self.port,"\n")
         print ("I will try a higher port")
         # store to user provideed port locally for later (in case 8080 fails)
         user_port = self.port
         self.port = 8080

         try:
             print("Launching HTTP server on ", self.host, ":",self.port)
             self.socket.bind((self.host, self.port))

         except Exception as e:
             print("ERROR: Failed to acquire sockets for ports ", user_port, " and
             print("Try running the Server in a privileged user mode.")
             self.shutdown()
             import sys
             sys.exit(1)

     print ("Server successfully acquired the socket with port:", self.port)
     print ("Press Ctrl+C to shut down the server and exit.")
     self._wait_for_connections()

 def shutdown(self):
     """ Shut down the server """
     try:
         print("Shutting down the server")
```

```
55:             s.socket.shutdown(socket.SHUT_RDWR)
56:
57:         except Exception as e:
58:             print("Warning: could not shut down the socket. Maybe it was already clos
59:
60:     def _gen_headers(self,  code):
61:         """ Generates HTTP response Headers. Ommits the first line! """
62:
63:         # determine response code
64:         h = ''
65:         if (code == 200):
66:            h = 'HTTP/1.1 200 OK\n'
67:         elif(code == 404):
68:            h = 'HTTP/1.1 404 Not Found\n'
69:
70:         # write further headers
71:         current_date = time.strftime("%a, %d %b %Y %H:%M:%S", time.localtime())
72:         h += 'Date: ' + current_date +'\n'
73:         h += 'Server: Simple-Python-HTTP-Server\n'
74:         h += 'Connection: close\n\n'  # signal that the conection wil be closed after
75:
76:         return h
77:
78:     def _wait_for_connections(self):
79:         """ Main loop awaiting connections """
80:         while True:
81:             print ("Awaiting New connection")
82:             self.socket.listen(3) # maximum number of queued connections
83:
84:             conn, addr = self.socket.accept()
85:             # conn - socket to client
86:             # addr - clients address
87:
88:             print("Got connection from:", addr)
89:
90:             data = conn.recv(1024) #receive data from client
91:             string = bytes.decode(data) #decode it to string
92:
93:             #determine request method  (HEAD and GET are supported)
94:             request_method = string.split(' ')[0]
95:             print ("Method: ", request_method)
96:             print ("Request body: ", string)
97:
98:             #if string[0:3] == 'GET':
99:             if (request_method == 'GET') | (request_method == 'HEAD'):
100:                #file_requested = string[4:]
101:
102:                # split on space "GET /file.html" -into-> ('GET','file.html',...)
103:                file_requested = string.split(' ')
```

```
104:                file_requested = file_requested[1] # get 2nd element
105:
106:                #Check for URL arguments. Disregard them
107:                file_requested = file_requested.split('?')[0]  # disregard anything a
108:
109:                if (file_requested == '/'):  # in case no file is specified by the bi
110:                    file_requested = '/index.html' # load index.html by default
111:
112:                file_requested = self.www_dir + file_requested
113:                print ("Serving web page [",file_requested,"]")
114:
115:                ## Load file content
116:                try:
117:                    file_handler = open(file_requested,'rb')
118:                    if (request_method == 'GET'):  #only read the file when GET
119:                        response_content = file_handler.read() # read file content
120:                    file_handler.close()
121:
122:                    response_headers = self._gen_headers( 200)
123:
124:                except Exception as e: #in case file was not found, generate 404 page
125:                    print ("Warning, file not found. Serving response code 404\n", e)
126:                    response_headers = self._gen_headers( 404)
127:
128:                    if (request_method == 'GET'):
129:                       response_content = b"<html><body><p>Error 404: File not found<
130:
131:                server_response =  response_headers.encode() # return headers for GET
132:                if (request_method == 'GET'):
133:                    server_response +=  response_content  # return additional conten
134:
135:                ##############################################
136:                #Our code to make the Javascript work
137:                ##############################################
138:                if request_method == 'GET' and string.count('?long') > 0:
139:                    user_input = input('What message do you want '
140:                                       'to send to the client?')
141:                    new_data = "<!DOCTYPE html><html><head><title>" \
142:                               "User Input</title></head><body>"\
143:                               + user_input \
144:                               + "</body></html>"
145:
146:                    conn.send(new_data.encode())
147:                else:
148:                    conn.send(server_response)
149:
150:
151:                print ("Closing connection with client")
152:                conn.close()
```

```
153:
154:          else:
155:              print("Unknown HTTP request method:", request_method)
156:
157: def graceful_shutdown(sig, dummy):
158:     """ This function shuts down the server. It's triggered
159:     by SIGINT signal """
160:     s.shutdown() #shut down the server
161:     import sys
162:     sys.exit(1)
163:
164: ###########################################################
165: # shut down on ctrl+c
166: signal.signal(signal.SIGINT, graceful_shutdown)
167:
168: print ("Starting web server")
169: s = Server(80)  # construct server object
170: s.activate_server() # aquire the socket
```

2 Web Crawler (10 Points)

Your task in this exercise is to "crawl" the `Simple English Wikipedia`. In order to execute this task, we provide you with a mirror of the Simple English Wikipedia at `141.26.208.82`.

You can start crawling from `http://141.26.208.82/articles/g/e/r/Germany.html` and you can use the `urllib` or `doGetRequest` function from the last week's assignment.

Given below is the strategy that you might adopt to complete this assignment:

1. Download `http://141.26.208.82/articles/g/e/r/Germany.html` and store the page on your file system.
2. Open the file in python and extract the local links. (Links within the same domain.)
3. Store the file to your file system.
4. Follow all the links and repeat steps 1 to 3.
5. Repeat step 4 until you have downloaded and saved all pages.

2.1 Hints:

- Before you start this exercise, please have a look at Exercise 3.
- Make really sure your crawler doesn't follow external urls to domains other than `http://141.26.208.82`. In that case you would start crawling the entire web
- Expect the crawler to run about 60 Minutes if you start it from the university network. From home your runtime will most certainly be even longer.
- It might be useful for you to have some output on the crawlers commandline depicting which URL is currently being fetched and how many URLs have been fetched so far and how many are currently on the queue.
- You can (but don't have to) make use of breadth-first search.
- It probably makes sense to take over the full paths from the pages of the Simple English Wikipedia and use the same folder structure when you save the html documents.
- You can (but you don't have to) speed up the crawler significantly if you use multithreading. However you should not use more than 10 threads in order for our mirror of Simple English Wikipedia to stay alive.

Answer: The urllib import statement only works in python 3.x

```
1: # assignment 5 task 2
2: # Andrea Mildes - mildes@uni-koblenz.de
```

```
3: # Sebastian Blei - sblei@uni-koblenz.de
4: # Johannes Kirchner - jkirchner@uni-koblenz.de
5: # Abdul Afghan - abdul.afghan@outlook.de
6:
7: # HTTP Requests
8: import socket
9:
10: # Parse the URLs und decode them
11: import urllib.parse
12:
13: # Required for the makedirs command which is used to create the folder structure
14: import os
15:
16: # Required for regular expressions
17: import re
18:
19: # Used for time-measuring (debug purposes)
20: import time
21:
22: # Globals
23: visited = set()
24: set_queue = set()
25: ext_links = 0
26: downloaded = 0
27: total_links = []
28:
29:
30: def recv_all(sock):
31:     data = b""
32:     part = None
33:     while part != b"":
34:         part = sock.recv(4096)
35:         data += part
36:     return data
37:
38:
39: # Parse the url in its components,
40: # create a socket and connect to a webserver on port 80.
41: # Send a GET Request and wait till all chunks arrived.
42: # If the next chunk is empty, continue.
43: def do_http_get_request(url):
44:     url = urllib.parse.urlparse(url)
45:     path = url.path
46:     if path == "":
47:         path = "/"
48:     host = url.netloc
49:     port = 80
50:     s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
51:
```

```
52:     s.connect((host, port))
53:     request = ""
54:     request += "GET " + path + " HTTP/1.0"
55:     request += "\r\n\r\n"
56:
57:     s.send(str.encode(request))
58:     response = recv_all(s)
59:
60:     if response:
61:         temp = response.split(b"\r\n\r\n", 1)
62:         header = temp[0]
63:         body = temp[1].strip()
64:
65:         return header, body, path
66:     else:
67:         print("No response :(")
68:         exit()
69:
70:
71: def save_to_file(body, path):
72:     global downloaded
73:     # remove first / in order to save the file in the directory of this .py file
74:     if path[0:1] == "/":
75:         path = path[1:]
76:
77:     # convert path to UTF-8
78:     path = urllib.parse.unquote_plus(path)
79:
80:     # create directories if not already exist
81:     if path.count("/") > 0:
82:         os.makedirs(os.path.dirname("dump/" + path), exist_ok=True)
83:
84:     file = open("dump/" + path, 'wb')
85:     file.write(body)
86:     file.close()
87:     downloaded += 1
88:
89:
90: def search_all_links(body, url):
91:     global ext_links, visited, set_queue, total_links
92:     q = set()
93:     count_external_links = 0
94:     count_internal_links = 0
95:
96:     url = urllib.parse.urlparse(url)
97:     pattern = re.compile("<a.+?href=[\"'](.+?)[\"'].*?>")
98:
99:     # Sometimes there is an Unicode Decode Error, so we simply catch it, log it ar
100:     try:
```

```
101:         links = pattern.findall(body.decode())
102:     except UnicodeDecodeError as ude:
103:         return -1
104:
105:     # Do stuff with all found links
106:     for i in range(0, len(links)):
107:         url_i = urllib.parse.urlparse(links[i])
108:
109:         # Ignore external links
110:         if url_i.netloc != '':
111:             if url_i.netloc != '141.26.208.82':
112:                 # Count all external links
113:                 count_external_links += 1
114:                 continue
115:             else:
116:                 break
117:
118:         # Count all internal links
119:         count_internal_links += 1
120:         count = links[i].count("../")
121:
122:         # Remove all ../ from the URL
123:         for j in range(0, count):
124:             index = links[i].find("/")
125:             links[i] = links[i][index+1:]
126:
127:         temp_path = url.path
128:         # count+1 because we have to remove the filename first :)
129:         for y in range(0, count+1):
130:             x = temp_path.rfind('/')
131:             temp_path = temp_path[:x]
132:
133:         # If the path has a leading / remove it
134:         if temp_path[0:1] == "/":
135:             temp_path = temp_path[1:]
136:
137:         result = url.scheme + "://" + url.netloc + "/" + temp_path + links[i]
138:
139:         # # Check if URL is already visited
140:         if result in visited or result in set_queue:
141:             continue
142:
143:         q.add(result)
144:
145:     total_links.append((count_internal_links + count_external_links, count_interna
146:     return q
147:
148:
149: def print_log(duration, decoding_error):
```

```
150:     duration_min = str(round(duration/60, 1)) + " min."
151:     duration_perc = "(" + str(round(((duration/1645)*100), 2)) + "%)"
152:     downloaded_perc = "(" + str(round(((downloaded/85322)*100), 1)) + "%)"
153:     file = open('log.txt', 'a')
154:     file.write("Duration: %-9s %-9s | Visited: %-8s | Queue: %-8s | Decoding-Erro
155:                (duration_min, duration_perc, len(visited), len(set_queue), decodi
156:     file.close()
157:
158:
159: def worker(starting_url):
160:     global visited, set_queue
161:     start_time = time.time()
162:     set_queue.add(starting_url)
163:     decoding_error = 0
164:
165:     c = 0
166:
167:     while len(set_queue) > 0:
168:         c += 1
169:
170:         # Logging
171:         if c % 1000 == 0:
172:             t = (time.time() - start_time)
173:             print_log(t, decoding_error)
174:
175:         current_url = set_queue.pop()
176:
177:         visited.add(current_url)
178:
179:         # Download current file
180:         header, response, path = do_http_get_request(current_url)
181:
182:         if header.count(b"200 OK") == 0:
183:             continue
184:
185:         save_to_file(response, path)
186:
187:         # If Unicode Error, log and continue
188:         link_set = search_all_links(response, current_url)
189:         if link_set != -1:
190:             set_queue.symmetric_difference_update(link_set)
191:         else:
192:             decoding_error += 1
193:             file = open('log.txt', 'a')
194:             file.write("\n ********** UTF8 DECODING ERROR ********** \n\n")
195:             file.close()
196:             continue
197:
198:     t = (time.time() - start_time)
```

```
199:     print_log(t, decoding_error)
200:     file = open('log.txt', 'a')
201:     file.write("\n ************************************ \n")
202:     file.write("\n\n ********** FINISHED CRAWLING ********** \n\n")
203:     file.write("\n ************************************ \n")
204:     file.close()
205:
206:     return
207:
208:
209: # returns all necessary information for task 3
210: def report_func():
211:     # Total amount of webpages
212:     # external and internal links per webbpage
213:     main_func()
214:
215:     # total links: [(total links, internal links, external links)]
216:     return len(visited), total_links
217:
218:
219: # call all functions
220: def main_func():
221:     # Create download folder
222:     os.makedirs(os.path.dirname("dump/"), exist_ok=True)
223:
224:     # Create a new file or overwrite the existing one
225:     file = open('log.txt', 'w')
226:     file.write("Start crawling ... \n")
227:     file.close()
228:     url = 'http://141.26.208.82/articles/g/e/r/Germany.html'
229:     worker(url)
230:
231:
232: if __name__ == '__main__':
233:     main_func()
```

3 Web Crawl Statistics (10 Points)

If you have successfully completed the first exercise of this assignment, then please provide the following details. You may have to tweak your code in the above exercise for some of the results.

3.1 Phase I

1. Total Number of *webpages* you found.
2. Total number of links that you encountered in the complete process of crawling.
3. Average and median number of links per web page.
4. Create a *histogram* showing the distribution of links on the crawled web pages. You can use a bin size of 5 and scale the axis from 0-150.

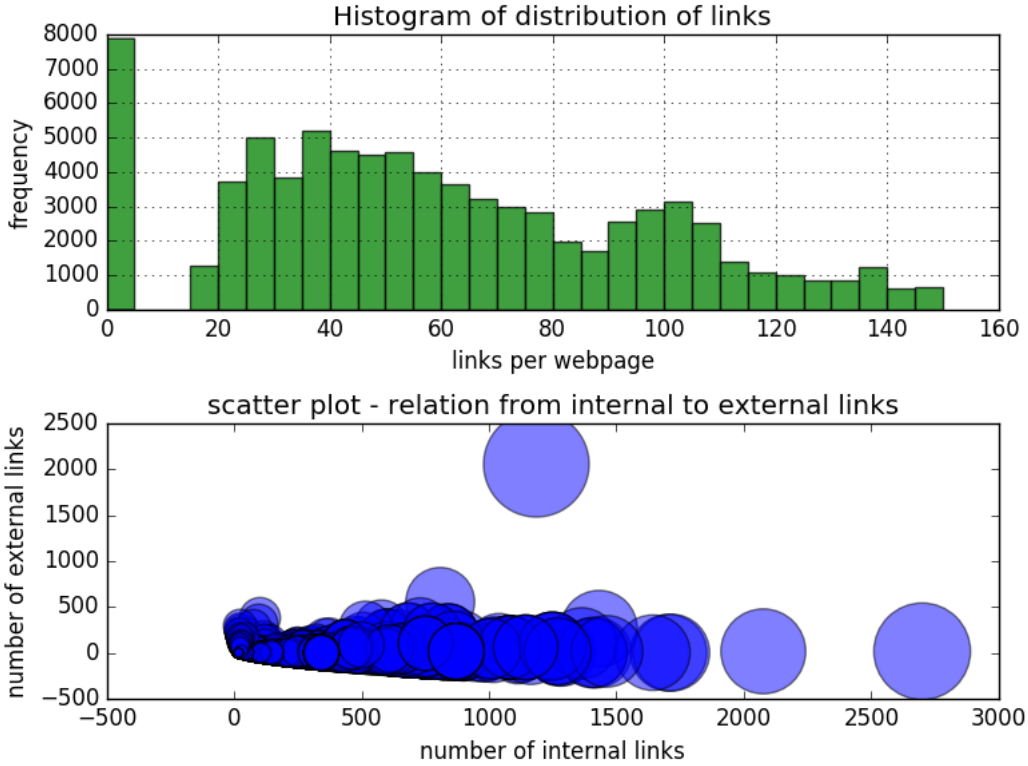
3.2 Phase II

1. For every page that you have downloaded, count the number of internal links and external links.
2. Provide a *scatter plot* with number of internal links on the X axis and number of external links on the Y axis.

Answer: Webpages by our definition are all internal pages we encountered, including 404 pages but excluding external pages.

```
# assignment 5 task 3
# Andrea Mildes - mildes@uni-koblenz.de
# Sebastian Blei - sblei@uni-koblenz.de
# Johannes Kirchner - jkirchner@uni-koblenz.de
# Abdul Afghan - abdul.afghan@outlook.de

import numpy as np
import matplotlib.pyplot as plt
import matplotlib.patches as lpatches
import web_crawler as crawl

visited, total_links = crawl.report_func()

all = 0
tmplist = []
for item in total_links:
        all += item[0]
        tmplist.append(item[0])

```

```
20: average = all/len(total_links)
21: sortlist = np.sort(tmplist)
22: medi = np.median(sortlist)
23:
24: print('total number of webpages: ' + str(visited))
25: print('total number of links encountered: ' + str(all))
26: print('average number of links per webpage: ' + str(average))
27: print('median number of links per webpage: ' + str(medi))
28:
29: linkCount = [link[0] for link in total_links]
30:
31: # histogram
32: plt.figure(1)
33: plt.subplot(211)
34: plt.xlabel('links per webpage')
35: plt.ylabel('frequency')
36: plt.title('Histogram of distribution of links')
37: #plt.axis([0, 150, 0, 1500])
38: plt.grid(True)
39: plt.hist(linkCount, bins=30, range=(0,150), facecolor='g', alpha=0.75)
40:
41: # scatterplot
42: plt.subplot(212)
43: plt.xlabel('number of internal links')
44: plt.ylabel('number of external links')
45: plt.title('scatter plot - relation from internal to external links')
46: for i in total_links:
47:         x = i[1]
48:         y = i[2]
49:         area = i[0]
50:         plt.scatter(x, y, s=area, alpha=0.5)
51: plt.tight_layout()
52: plt.show()
```


Histogram of distribution of links
frequency
links per webpage
scatter plot - relation from internal to external links
number of external links
number of internal links

Important Notes

Submission

- Solutions have to be checked into the github repository. Use the directory name `groupname/assignment5/` in your group's repository.
- The name of the group and the names of all participating students must be listed on each submission.
- Solution format: all solutions as *one* PDF document. Programming code has to be submitted as Python code to the github repository. Upload *all* `.py` files of your program! Use `UTF-8` as the file encoding. *Other encodings will not be taken into account!*
- Check that your code compiles without errors.
- Make sure your code is formatted to be easy to read.
  - Make sure you code has consistent indentation.
  - Make sure you comment and document your code adequately in English.
  - Choose consistent and intuitive names for your identifiers.
- Do *not* use any accents, spaces or special characters in your filenames.

Acknowledgment

This latex template was created by Lukas Schmelzeisen for the tutorials of "Web Information Retrieval".

LaTeX

Currently the code can only be build using LuaLaTeX, so make sure you have that installed. If on Overleaf, there's an error, go to settings and change the LaTeXengine to `LuaLaTeX`.