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.

Group Name : Yankee
Candidates :

1. Sabin Bhattarai - 216203590
   sbhattarai@uni-koblenz.de
2. Biplov K.C. - 216203865
   biplov@uni-koblenz.de
3. Syed Salman Ali. - 216203923
   salmanali@uni-koblenz.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 1

Please note we have provided a file index.html inside www folder during submission./

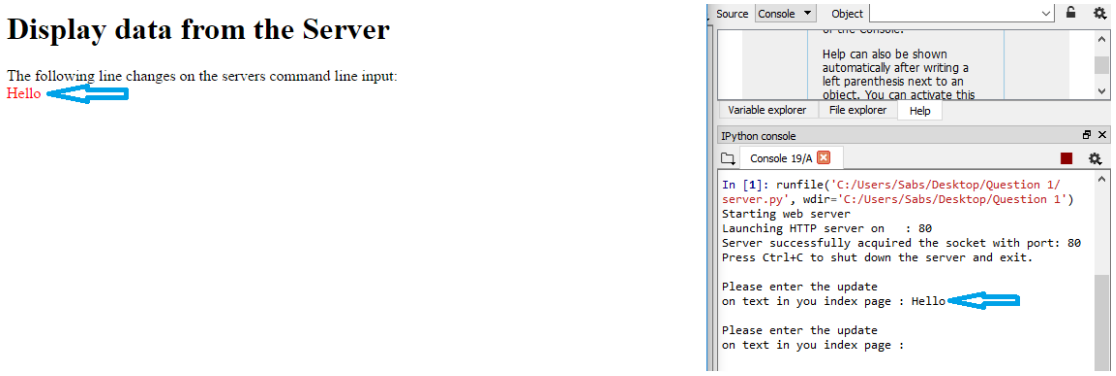

Figure 1: Shows the update when text typed in command line from server end

Below we show the Server Side code downloaded as told and edited/

```
#ServerSide Code
#!/usr/bin/python

import socket  # Networking support
import signal  # Signal support (server shutdown on signal receive)
import time    # Current time
import sys
from threading import Thread


inputStringFromCommandGlobal = "This will be replaced by messages from the server"

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)
     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 8080. ")
             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)
```

```
49:       print ("Press Ctrl+C to shut down the server and exit.")
50:       self._wait_for_connections()
51:
52:   def shutdown(self):
53:       """ Shut down the server """
54:       try:
55:           print("Shutting down the server")
56:           s.socket.shutdown(socket.SHUT_RDWR)
57:
58:       except Exception as e:
59:           print("Warning: could not shut down the socket. Maybe it was \
60:                                                   already closed?",e)
61:
62:   def _gen_headers(self,  code):
63:       """ Generates HTTP response Headers. Ommits the first line! """
64:
65:       # determine response code
66:       h = ''
67:       if (code == 200):
68:          h = 'HTTP/1.1 200 OK\n'
69:       elif(code == 404):
70:          h = 'HTTP/1.1 404 Not Found\n'
71:
72:       # write further headers
73:       current_date = time.strftime("%a, %d %b %Y %H:%M:%S", time.localtime())
74:       h += 'Date: ' + current_date +'\n'
75:       h += 'Server: Simple-Python-HTTP-Server\n'
76:       # signal that the conection wil be closed after complting the request
77:       h += 'Connection: close\n\n'
78:
79:       return h
80:
81:   def _wait_for_connections(self):
82:       """ Main loop awaiting connections """
83:       while True:
84:           #print ("Awaiting New connection")
85:           self.socket.listen(3) # maximum number of queued connections
86:
87:           conn, addr = self.socket.accept()
88:           # conn - socket to client
89:           # addr - clients address
90:
91:           #print("Got connection from:", addr)
92:
93:           data = conn.recv(1024) #receive data from client
94:           string = bytes.decode(data) #decode it to string
95:
96:           #determine request method  (HEAD and GET are supported)
97:           request_method = string.split(' ')[0]
```

```
98:         #print ("Method: ", request_method)
99:         #print ("Request body: ", string)
100:
101:        #if string[0:3] == 'GET':
102:        if (request_method == 'GET') | (request_method == 'HEAD'):
103:            #file_requested = string[4:]
104:
105:            # split on space "GET /file.html" -into-> ('GET','file.html',...)
106:            file_requested = string.split(' ')
107:            file_requested = file_requested[1] # get 2nd element
108:
109:            #Check for URL arguments. Disregard them
110:            file_requested = file_requested.split('?')[0]  # disregard anything a
111:
112:            if (file_requested == '/'):  # in case no file is specified by the bi
113:                file_requested = '/index.html' # load index.html by default
114:
115:            file_requested = self.www_dir + file_requested
116:            #print ("Serving web page [",file_requested,"]")
117:
118:            ## Load file content
119:            try:
120:                file_handler = open(file_requested,'rb')
121:                if (request_method == 'GET'):  #only read the file when GET
122:                    response_content = file_handler.read() # read file content
123:                file_handler.close()
124:
125:                response_headers = self._gen_headers( 200)
126:
127:            except Exception as e: #in case file was not found, generate 404 page
128:                if(file_requested == 'www/GETINPUT' and request_method == 'GET'):
129:                    response_headers = self._gen_headers( 200)
130:                    #print ("oh its a input file")
131:                    response_content = inputStringFromCommandGlobal.\
132:                                                            encode('utf-8')
133:                else:
134:                    response_headers = self._gen_headers( 404)
135:                    response_content = b"<html><body><p>Error 404: File not \
136:                                found</p><p>Python HTTP server</p></body></html>"
137:
138:            # return headers for GET and HEAD
139:            server_response =  response_headers.encode()
140:            if (request_method == 'GET'):
141:                #return additional conten for GET only
142:                server_response +=  response_content
143:
144:            conn.send(server_response)
145:            conn.close()
146:
```

```
147:
148: def graceful_shutdown(sig, dummy):
149:     """ This function shuts down the server. It's triggered
150:     by SIGINT signal """
151:     s.shutdown() #shut down the server
152:     sys.exit(1)
153:
154:
155: # shut down on ctrl+c
156: signal.signal(signal.SIGINT, graceful_shutdown)
157:
158: print ("Starting web server")
159: s = Server(80) # construct server object
160:
161: def runServer():
162:     s.activate_server() # aquire the socket
163:
164: #Starting a server thread
165: t1 = Thread(target=runServer)
166: t1.start()
167:
168: #Main Thread running for command line inputs.
169: while True:
170:     inputStringFromCommandGlobal = input("Please enter the update \
171:                                          on text in you index page : ")
```

The following is the index.html

```
#index.html
<html>
<head>
    <title>
        Abusing the HTTP protocol - Example
    </title>
</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>


    <script language="javascript">
        function checkUpdates() {
            var XMLHttpRequestObject = false;
            if (window.XMLHttpRequest) {
                XMLHttpRequestObject = new XMLHttpRequest();
            } else if (window.ActiveXObject) {
                XMLHttpRequestObject = new ActiveXObject("Microsoft.XMLHTTP");
            }
            if (XMLHttpRequestObject) {
                var obj = document.getElementById("response");
                XMLHttpRequestObject.open("GET", "GETINPUT");
                XMLHttpRequestObject.onreadystatechange = function () {
                    if (XMLHttpRequestObject.readyState == 4 &&
                            XMLHttpRequestObject.status == 200) {
                        obj.innerHTML = XMLHttpRequestObject.responseText;
                    }

                    checkUpdates();
                }
                XMLHttpRequestObject.send()
            }
        }
        checkUpdates()

    </script>

</body>
</html>
```

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.

The following is the crawl.py

```
#crawl.py
import re
import socket
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from urllib.parse import urlparse
from urllib.parse import urljoin

ENCODING = 'utf-8'
url_to_process = {'http://141.26.208.82/articles/g/e/r/Germany.html' : 0}
url_visited = {}
url_external_internal = {}
url_totalLink = {}


#Helper function that handles socket connection and also gives response in bytes
def _clientConnectAndResponse(url, timeout=10, receive_buffer=8096):
    parsed = urlparse(url)
    try:
        host, port = parsed.netloc.split(':')
    except ValueError:
        host, port = parsed.netloc, 80

    try:
        clientSocket = socket.create_connection((host, port), timeout)
    except socket.timeout:
        raise


    method  = 'GET %s HTTP/1.0\r\n\r\n' % parsed.path
    clientSocket.sendall(bytes(method, ENCODING))

    response = [clientSocket.recv(receive_buffer)]
    while response[-1]:
        response.append(clientSocket.recv(receive_buffer))

    responseInBytes = b''.join(r for r in response)
    return responseInBytes


# Connects to the server and receives response. writes to a file,
# returns filename , returns None if not 200 OK
def _downloadHTMLToFile(url,content_typeLength=14 ):
    #Connecting to Server and obtaining responses in Byte format
    try:
        responseInBytes = _clientConnectAndResponse(url)
```

```
49:     except:
50:         return None
51:
52:     headerBytes = responseInBytes[:responseInBytes.find(b'\r\n\r\n')]
53:     remainingBytes = responseInBytes[responseInBytes.find(b'\r\n\r\n'):]\
54:                                      .strip(b'\r\n\r\n')
55:
56:     #Here we check if HTTP responses with 200 OK
57:     if (headerBytes.find(b'\r\n')) > 0 :
58:         if headerBytes[:headerBytes.find(b'\r\n')].decode(ENCODING).lower()\
59:                         != "http/1.1 200 ok":
60:             #print ("Cannot determine if HTTP 200 OK. Further Processing halts")
61:             return None
62:
63:     # Check if url has a path. This is to extract filename for writing to file.
64:     if (urlparse(url).path == ''):
65:         filename =  url.split('#')[-1]
66:     else:
67:         filename =  urlparse(url).path.split('/')[-1]
68:
69:     # Write to a file with name given by filename variable
70:     try:
71:         with open(filename, 'wb') as htmlToFile:
72:             htmlToFile.write(remainingBytes)
73:             htmlToFile.close()
74:         return filename
75:     except IOError as e:
76:         return None
77:
78:
79: # Helper function used by parseHTMLForLinks to check if url is internal
80: def _is_internal(urlMain, urlToCheck):
81:   return urlparse(urlMain).scheme == urlparse(urlToCheck).scheme and\
82:                       urlparse(urlMain).netloc == urlparse(urlToCheck).netloc
83:
84:
85: # Helper function that returns the front part of the url that is to be
86: # appeneded to relative urls for valid image construction
87: def _getPartToAppendToRelativeLinkURL(url):
88:     parsed = urlparse(url)
89:     try:
90:         host, port = parsed.netloc.split(':')
91:     except ValueError:
92:         host = parsed.netloc
93:     return parsed.scheme +'://'+ host
94:
95:
96: # obtaines a dictionary of internal and external links for a given url
97: # Done by opening the file and extracting all links. Manages unique links
```

```
 98: def _parseHTMLForLinks(filename , url, content_typeLength=14):
 99:     externalLinks = set()
100:     internalLinks = set()
101:     allLinks = {}
102:
103:     with open(filename , encoding = ENCODING) as html:
104:         try:
105:             content = html.read()
106:         except:
107:             allLinks['internal'] = internalLinks
108:             allLinks['external'] = externalLinks
109:             return allLinks
110:         pat = re.compile ('<a href="?\'?([^"\'>]*)')
111:         linksInAHref = pat.findall(content)
112:
113:     toAppendRelativeLinkURL = _getPartToAppendToRelativeLinkURL(url)
114:     for linkInAHref in linksInAHref:
115:         if(linkInAHref.find('://') < 0):
116:
117:             #We have a relative link. So need to modify link
118:             link = urljoin(toAppendRelativeLinkURL, linkInAHref)
119:
120:             internalLinks.update([link])
121:         else:
122:             if _is_internal(url,linkInAHref ):
123:                 internalLinks.update([linkInAHref])
124:             else:
125:                 externalLinks.update([linkInAHref])
126:
127:     allLinks['internal'] = internalLinks
128:     allLinks['external'] = externalLinks
129:     return allLinks
130:
131: # Used by process_UrlBFS to obtain all the link for a given url. Uses
132: # two helper function to download html to file and also parse html to get links
133: def getAllLinksForThisURL(url):
134:     #making sure all links obtained inside from the url is unique
135:     links = {'internal' : {} , 'external':{}}
136:
137:     fileLocation = _downloadHTMLToFile(url)
138:     if fileLocation is not None:
139:         links = _parseHTMLForLinks(fileLocation,url)
140:
141:     return links
142:
143: # Here we obtain the URL from global variable and start the URL crawling.
144: def process_urlBFS():
145:     totalLinks = 0
146:     totalPages = 0
```

```
147:     while(url_to_process):
148:         key , value  = url_to_process.popitem()
149:         # Lets process the url to get more url links.
150:         # Also download the html file for this url. Returns all link info
151:         receivedInternalExternalLinks = getAllLinksForThisURL(key)
152:
153:         url_totalLink[key] = len(receivedInternalExternalLinks['internal']) +\
154:                                  len(receivedInternalExternalLinks['external'])
155:
156:         #This will be used to calculate pages as we take 200 ok link as pages.
157:         url_visited[key] = url_totalLink[key]
158:
159:         #key as url associated to values tuple i.e (internal, external)
160:         url_external_internal[key] = (len(receivedInternalExternalLinks['internal
161:                                             ]),len(receivedInternalExternalLinks\
162:                                             ['external']))
163:
164:         totalLinks = totalLinks + len(receivedInternalExternalLinks['internal'])\
165:                                           + len(receivedInternalExternalLinks\
166:                                           ['external'])
167:         allInternalReceivedURLLinks = receivedInternalExternalLinks['internal']
168:
169:
170:         for receivedURLLink in allInternalReceivedURLLinks:
171:                 # check if contains in url_visited. Add to process queue
172:                 if (receivedURLLink not in url_visited):
173:                     url_to_process[receivedURLLink] = 0
174:
175:     totalPages = len([x for x in url_visited.values() if x != 0])
176:     #badPages404 = len([x for x in url_visited.values() if x == 0])
177:     print ("\nPhase I")
178:     print("\nTotal Number of webpages we found : " , totalPages)
179:     print("\nTotal Number of links we encountered :" , totalLinks)
180:
181:     average = totalLinks / totalPages
182:     print("\nAverage number of links per web page : ", average )
183:     print ("\nMedian number of links per web page : ", \
184:                                               calculateMedian(url_totalLink))
185:     drawHistogram(url_totalLink)
186:     print ("\nPhase II")
187:     drawScatterDiagram(url_external_internal)
188:
189: def calculateMedian(url_totalLink):
190:     medianValue = (np.median(list(url_totalLink.values())))
191:     return medianValue
192:
193: def drawHistogram(url_totalLink):
194:     x_1 = np.array(list(url_totalLink.values()))
195:     plt.figure(figsize=(7,5))
```

```
196:     plt.hist(x_1 , bins= range(min(x_1), 150, 5))
197:     plt.xlabel("Number of Links per page (Each bar i.e. 0-5, 5-10, ...)")
198:     plt.ylabel("Frequencies")
199:     plt.show()
200:
201: def drawScatterDiagram(url_external_internal):
202:     #preparing data for plotting
203:     tuple_internal_external = list(url_external_internal.values())
204:     internal = [x[0] for x in tuple_internal_external]
205:     external = [x[1] for x in tuple_internal_external]
206:     #print ("mero internal lsit : " , internal )
207:     #print ("mero external lsit : " , external )
208:     df1 = pd.DataFrame(index=range(len(internal)), columns=["URL" , \
209:                                             "Number Of Internal Links",\
210:                                             "Number Of External Links"])
211:     df1["URL"] = url_external_internal.keys()
212:     df1["Number Of Internal Links"] = internal
213:     df1["Number Of External Links"] = external
214:     print(df1)
215:     sns.lmplot(x="Number Of Internal Links", y="Number Of External Links"\
216:                                                     , data=df1, fit_reg=False)
217:
218: def mainFunction():
219:     process_urlBFS()
220:     print('done')
221:
222:
223:
224: if __name__ == "__main__":
225:     mainFunction()
```

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.

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`.