Introduction to Web Science

Assignment 5

PD Dr. Matthias Thimm
thimm@uni-koblenz.de

Ipek Baris Schlicht
ibaris@uni-koblenz.de

Kenneth Skiba
kennethskiba@uni-koblenz.de

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

Submission until: 15.12.2020, CEST 23:59

Team: Bravo
Members:
Gaurav Kumar (220200656)
Pavithree Shetty (220200661)
Nisha Sharma (220202359)

1 Analysis of Simple English Wikipedia 60 Points

1.1 Crawling

Your task in this exercise is to crawl the Simple English Wikipedia. In order to execute this task, please first follow the steps in **Installation**.

Installation

In order to solve the task, you are required to download Simple English Wiki Dump `wikipedia_en_simple_all_nopic_2020-10.zim` and then set up `dockerized Kiwix Server` for hosting the Wiki locally. Please follow these installation steps:

1. Install a `Docker` engine (https://docs.docker.com/engine/install/) if your computer does not have one.
2. Download the Simple English Wiki 2020 (`wikipedia_en_simple_all_nopic_2020-10.zim`) from the website (https://ftp.fau.de/kiwix/zim/wikipedia/). The size of the file is 481M and it does not contain pictures.
3. Clone the repository (https://github.com/isspek/dockerized-kiwix-server) where you can find the installation scripts for installing Wiki Server using `Docker`. Follow the installation steps in the repository.
4. You are supposed to see Simple English Wikipedia at `localhost:8080`, when the installation is successfully complete.

Task

You can start crawling from http://localhost:8080/wikipedia_en_simple_all_nopic_2020-10/A/COVID-19_pandemic as seed set and you can use the libraries `beautifulsoup`, `request`, `urllib`, `pandas`, `re`, `numpy`, `collections`, `logging`, `pickle`, `sys`, `dataclasses`, `matplotlib`.

A simple crawler has minimum following units:

1. A fetcher for retrieving urls.
2. A parser for extracting contents.
3. A filter to handle duplicate or already fetched urls.

Additionally you are expected to handle any errors such as memory leak, http errors.

Hints

- Before you start this exercise, please have a look at 1.3.
- Make really sure your crawler does not follow external urls to domains other than http://localhost:8080/wikipedia_en_simple_all_nopic_2020-10/A/.
- Crawling all pages might take time, be patient.

- It might be useful for you to have some output on the crawlers command line 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.
- You can (but you don't have to) speed up the crawler significantly if you use multithreading.

1.2 Limitations

Briefly explain the potential limitations of your crawler (max 200 words). **Hint:** Think use cases for different websites when you use your crawler.

Solution

- **Difficult to Analyze:** For anybody, the extracting process is confusing to read.
- **Data Analysis:** The links that has been extracted will first need to be treated so that they can be easily understood. In some cases, this might take a long time and a lot of energy to complete.
- **Time:** It is normal for new data extraction applications to take some time in the beginning to become familiar with the core application and need to adjust to the scrapping language.
- **Speed and protection policies:** Most web scrapping services are slower than API calls and another problem is the websites that do not allow screen scraping. In such cases, web scrapping services are rendered useless.

1.3 Web Crawl Statistics

If you have successfully completed the first exercise of this assignment, then please provide the following details:

1. Total number of links that you fetched in the complete process of crawling.
2. Top 10 Wiki pages that have been linked more than once. Print them with their counts and save them as bar plot figure.
3. Top 10 external pages (pages with a domain which is not `wikipedia_en_simple_all_nopic_2020-10`) that have been linked more than once. Print them with their counts and save them as bar plot figure.
4. For every Wiki page that you have read, count the unique number of internal links and external links. Then provide a histogram for internal links and external links. You can use a bin size of 5 and limit ranges between 0 and 50. Make sure that there is a title indicating mean and sigma values. Save both histogram in separate files.

5. Additionally, save print logs of your task as a file with `.log` extension.

```
Total number of links that you fetched in the complete process of crawling: 45905
```

Figure 1: Total Crawl count

```
Top 10 Wiki pages that have been linked more than once
World_Health_Organization  : 12
List_of_states_with_limited_recognition  : 11
Wuhan  : 6
#cite_note-JHU_ticker-6  : 6
COVID-19_pandemic_in_Georgia_(country)  : 5
Coronavirus_disease_2019  : 4
Severe_acute_respiratory_syndrome_coronavirus_2  : 4
Hubei  : 4
COVID-19_pandemic_in_Azerbaijan  : 4
#cite_note-LancetI-432  : 4
```

Figure 2: Wiki top 10 pages

```
Top 10 external pages (pages with a domain which is not wikipedia_en_simple_all_nopic_2020-10) that have been linked more than
https://www.who.int/news-room/q-a-detail/q-a-coronaviruses  : 3
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7323513  : 2
https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6  : 2
https://www.who.int/emergencies/diseases/novel-coronavirus-2019  : 2
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102572  : 2
https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html  : 2
https://www.rivm.nl/coronavirus-covid-19/actueel  : 2
http://moh.gov.so/en/covid19/  : 2
https://covid19.who.int/  : 2
https://www.ssi.dk/sygdomme-beredskab-og-forskning/sygdomsovervaagning/c/covid19-overvaagning  : 2
```

Figure 3: External top 10 pages

```
The Mean is 631.2857142857143 and Standard deviation is 529.9688092477803 of Internal Links
The Mean is 284.5416666666667 and Standard deviation is 240.52087213078943 of External Links
```

Figure 4: Mean and Standard deviation

web_crawler.log

The file size (5.35 MB) exceeds configured limit (2.56 MB). Code insight features are not available.

```
92257  INFO:root:Sub Link: Wuhan
92258  INFO:root:Sub Link: https://simple.wikipedia.org/wiki/?title=COVID-19_pandemic&oldid=7130634
92259  INFO:root:Sub Link: https://creativecommons.org/licenses/by-sa/4.0/
92260  INFO:root:Total number of links that you fetched in the complete process of crawling: 45697
92261  INFO:root:Top 10 Wiki pages that have been linked more than once
92262  INFO:root:World_Health_Organization : 12
92263  INFO:root:List_of_states_with_limited_recognition : 11
92264  INFO:root:Wuhan : 6
92265  INFO:root:#cite_note-JHU_ticker-6 : 6
92266  INFO:root:COVID-19_pandemic_in_Georgia_(country) : 5
92267  INFO:root:Coronavirus_disease_2019 : 4
92268  INFO:root:Severe_acute_respiratory_syndrome_coronavirus_2 : 4
92269  INFO:root:Hubei : 4
92270  INFO:root:COVID-19_pandemic_in_Azerbaijan : 4
92271  INFO:root:#cite_note-LancetI-432 : 4
92272  INFO:root:Top 10 external pages (pages with a domain which is not wikipedia_en_simple_all_nopic_2020-10) that have b
92273  INFO:root:https://www.who.int/news-room/q-a-detail/q-a-coronaviruses : 3
92274  INFO:root:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7323513 : 2
92275  INFO:root:https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 : 2
92276  INFO:root:https://www.who.int/emergencies/diseases/novel-coronavirus-2019 : 2
92277  INFO:root:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102572 : 2
92278  INFO:root:https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html : 2
92279  INFO:root:https://www.rivm.nl/coronavirus-covid-19/actueel : 2
92280  INFO:root:http://moh.gov.so/en/covid19/ : 2
92281  INFO:root:https://covid19.who.int/ : 2
92282  INFO:root:https://www.ssi.dk/sygdomme-beredskab-og-forskning/sygdomsovervaagning/c/covid19-overvaagning : 2
92283  INFO:root:The Mean is 627.8367346938776 and Standard deviation is 533.368438991085 of Internal Links
92284  INFO:root:The Mean is 283.7291666666667 and Standard deviation is 241.36605495106159 of External Links
92285  INFO:root:Top 10 Wiki pages that have been linked more than once
92286  INFO:root:World_Health_Organization : 12
92287  INFO:root:List_of_states_with_limited_recognition : 11
92288  INFO:root:Wuhan : 6
92289  INFO:root:#cite_note-JHU_ticker-6 : 6
92290  INFO:root:COVID-19_pandemic_in_Georgia_(country) : 5
92291  INFO:root:Coronavirus_disease_2019 : 4
92292  INFO:root:Severe_acute_respiratory_syndrome_coronavirus_2 : 4
92293  INFO:root:Hubei : 4
92294  INFO:root:COVID-19_pandemic_in_Azerbaijan : 4
92295  INFO:root:#cite_note-LancetI-432 : 4
92296  INFO:root:Top 10 Wiki pages that have been linked more than once
92297  INFO:root:World_Health_Organization : 12
92298  INFO:root:List_of_states_with_limited_recognition : 11
92299  INFO:root:Wuhan : 6
92300  INFO:root:#cite_note-JHU_ticker-6 : 6
92301  INFO:root:COVID-19_pandemic_in_Georgia_(country) : 5
```

Figure 5: Log file

2 Questions 20 Points

Answer the following questions with your own words.

1. Why do we have to run probabilistical model more then once? (max 200 words)

   a) Random variables and probability distributions are incorporated into the model of an event or phenomenon in probabilistic models, giving a probability distribution as a solution.

   b) Since Probabilistic models involve random process, we will not get the exact same results everytime we run the model.

   c) By running it just once it can produce possible outliers.

   d) So we run the probabilistic model several times to get the most optimum result.

   e) Also,By running it more than once we get statistical stability, with less fluctuations in the results.

   f) In addition to this, its advisable to run every scientific experiment more than once to avoid mistakes.

2. Given the vector $v = [5, 6, 8, 12, 13, 100]$ calculate the mean and the median

   a) Solution is shown in figure 1

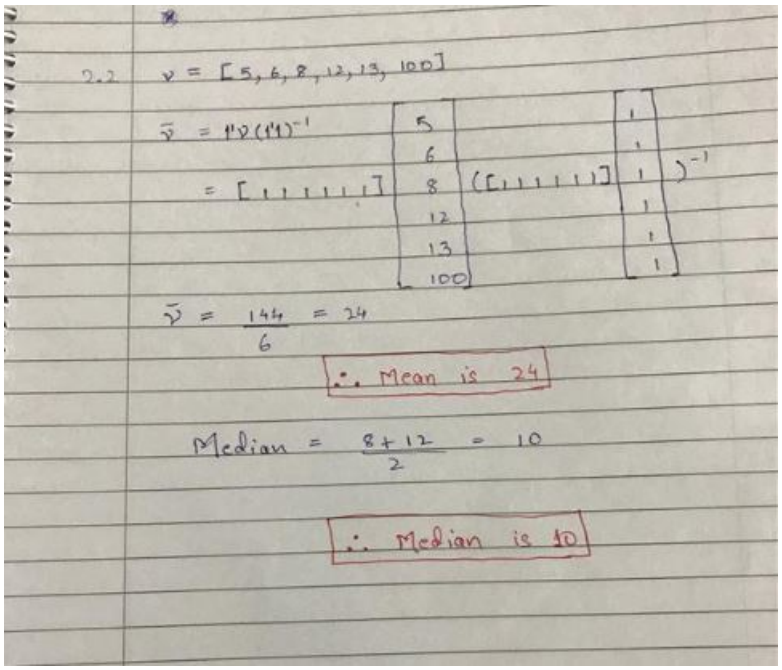

Figure 6: Solution

3. Given the following hypothesis say whether they are falsifiable, and explain why.
   - *All swans are white.*
     a) This hypothesis is falsifiable.
     b) We can falsify the above hypothesis by finding a swan which is not white.
   - *She will go running, when it rains or not.*
     a) The above hypothesis is falsifiable.
     b) We can falsify the above hypothesis by finding an instance where she doesn't go running, when it rains or not.(Maybe a day when she is sick and is unable to go running)
   - *The grass is wet, so it must have rained.*
     a) The above hypothesis is falsifiable.
     b) We can falsify the above hypothesis by finding an occasion when it doesn't rain but the grass is still wet (maybe the grass is wet due to water from sprinkler).
4. Given the plot in Figure 7[1] formulate **one** sentence that describes the plot.
   a) There is significant rise in the number of internet hosts with less than 10 internet hosts in the year 1970 to over 1G hosts by 2018.

[1]Based on https://en.wikipedia.org/wiki/File:Internet_Hosts_Count_log.svg

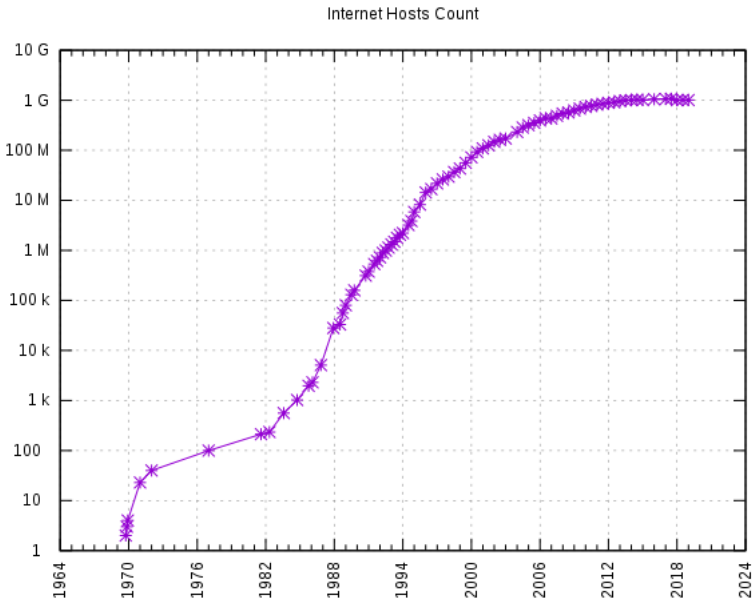

Figure 7: Number of Internet Hosts Counts