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

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.

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.

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 incor[porated 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

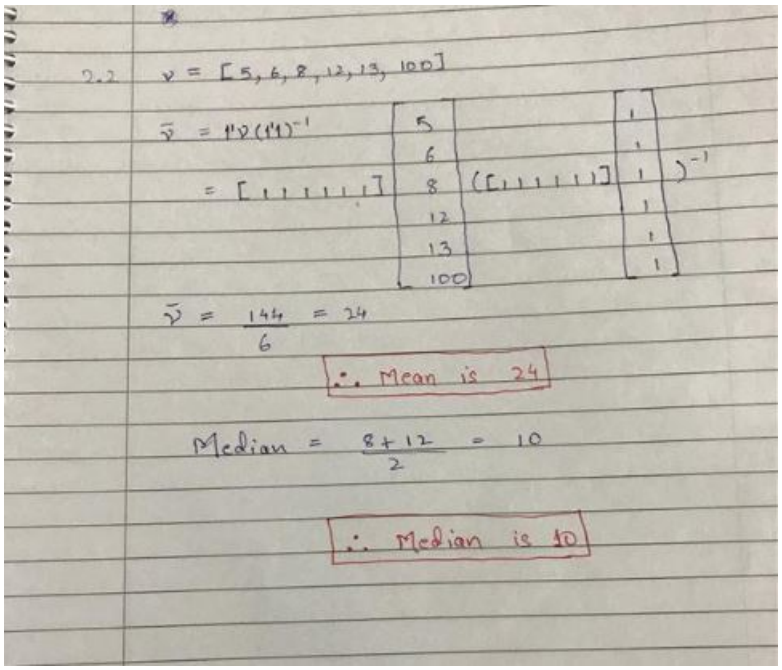

Figure 1: Solution

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 findind a single 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 2[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

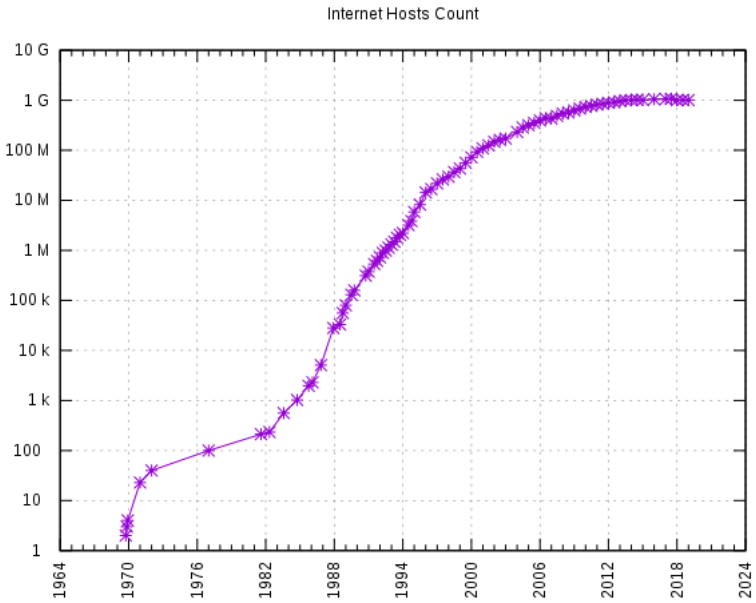

Figure 2: Number of Internet Hosts Counts