Crawling and Analysis of Gaming-Related Websites

Implementation Details

The implementation of our crawler starts from 4 gaming-related websites:

1. https://kotaku.com
2. https://sg.yahoo.com/topics/sg-gaming
3. https://www.channelnewsasia.com/topic/gaming
4. https://sea.ign.com

We set the upper limit of websites to be visited at 4096. This means that the execution of our crawler will automatically stop once we have visited >= 4096 URLs, including those that are blocked (based on robots.txt).

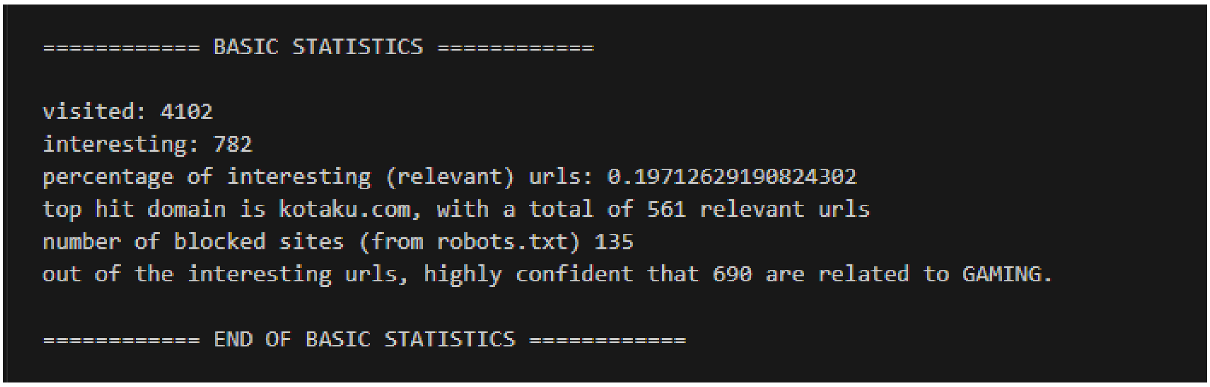
To determine whether a website is related to gaming, we have a set of keywords defined in *keywords.txt*. This was manually created based on some research on what are popular keywords related to video games. Then, we filter URLs based on whether there are any matches with the keywords and triage them as “interesting URLs”.

At the end of the crawling process, we re-visit the interesting URLs and generate a summary of the HTML page, using *newspaper3k* as the HTML content parser, and *spaCy* (NLP library) to generate a summary of the HTML content. Then, we take the top 5 sentences with the highest sentence strength and check if any one of them contains gaming-related keywords. If it does, then we classify this URL with higher confidence that it is gaming-related. The next section details our findings.

Findings

We have included all relevant files that have been used for the analysis under the *data* folder. **This comes in the form of a *.db* file, and 5 JSON files**. Note that the demo video only showcases very briefly the functionalities of our program.

From running the crawler with an upper limit of 4096 URLs to visit each time, we found the following (output generated by the program):



```
============ BASIC STATISTICS ============

visited: 4102
interesting: 782
percentage of interesting (relevant) urls: 0.19712629190824302
top hit domain is kotaku.com, with a total of 561 relevant urls
number of blocked sites (from robots.txt) 135
out of the interesting urls, highly confident that 690 are related to GAMING.

============ END OF BASIC STATISTICS ============
```

More than 4096 URLs are visited as threads will continue to run even after we exit of the loop.

The **number of URLs attempted to be visited were 4102**. Out of which**, 135 websites were blocked** based on rules defined in *robots.txt*. Hence, a total of **3967 rows (URLs) are in the *a4-table.db* file**. This database file contains the URL name, response time, IP address and geolocation of that particular URL.

We found that there were a total of **782 potentially interesting URLs**, which accounts for ~19.7% of the URLs visited. The domain with the **greatest number of URLs** that were **interesting is https://kotaku.com with a total of 561 relevant URLs.**

Based on the methodology described in our implementation, we **are highly confident that 690 out of 782 (approx. 88.2%) of the interesting URLs are related to gaming.**

Lastly, there are some websites found by the crawler during the run that have **yet to be visited**. This can be found in the *remaining.json* file. There are a total of **853 URLs**, bringing our total URL count (visited + not visited) to 4102 + 853 = 4955. **The websites not visited account for approximately 17.2% of all URLs found.**

Since we have yet to visit them, we did not add the details into our database file. We also decided not to triage them.

The pie charts below shows a breakdown of each and every single one of the details we have recorded, with annotations provided.

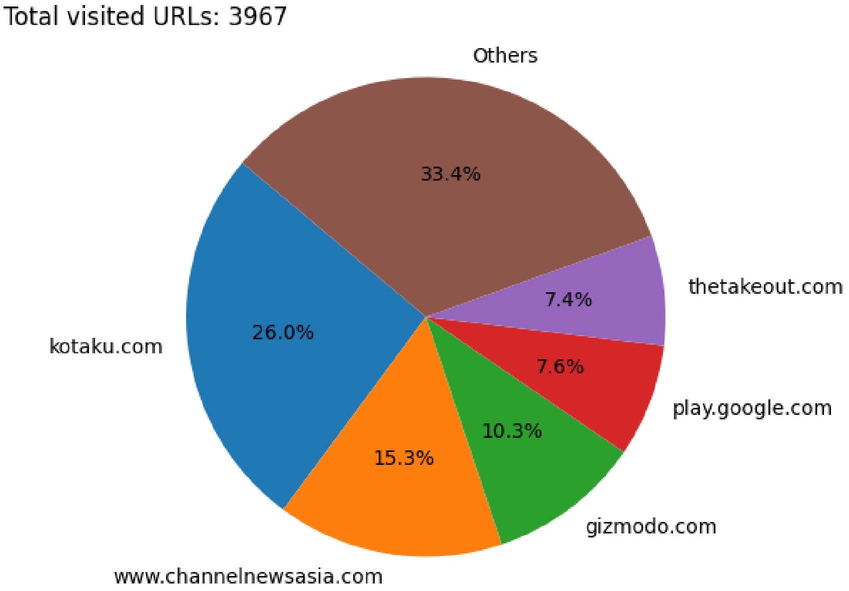


Figure 1: Breakdown (by percentage and domain) of visited URLs

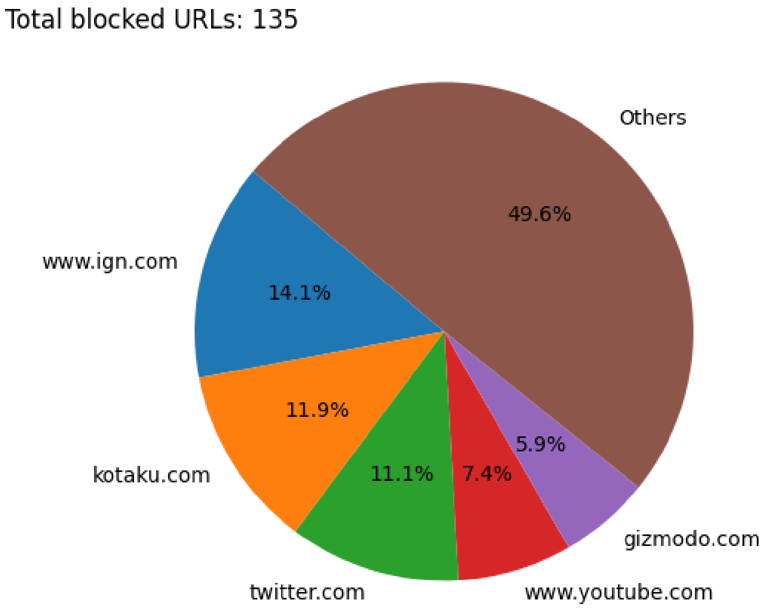


Figure 2: Breakdown (by percentage and domain) of blocked URLs

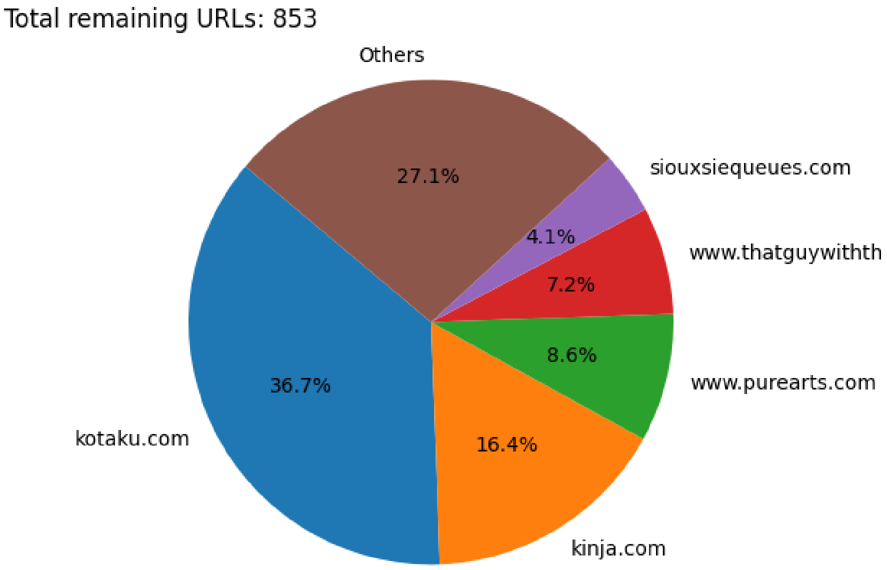


Figure 3: Breakdown (by percentage and domain) of remaining URLs to visit

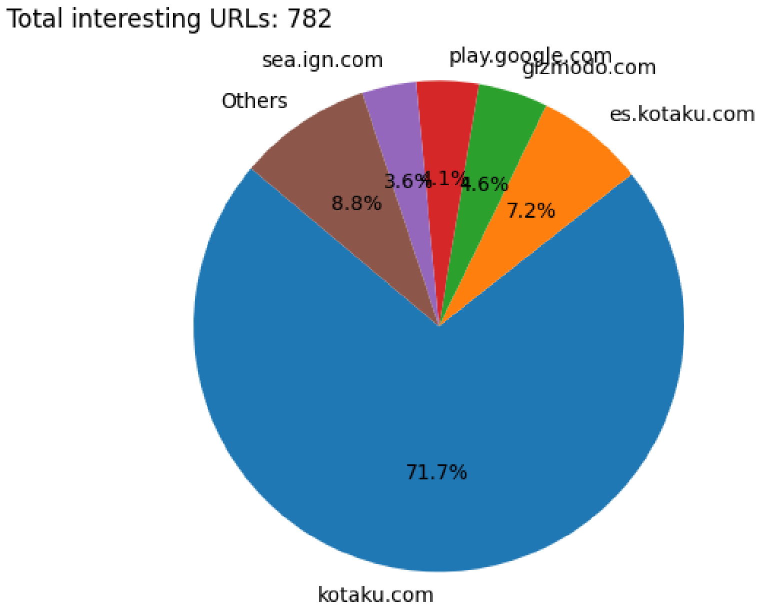


Figure 4: Breakdown (by percentage and domain) of interesting URLs visited

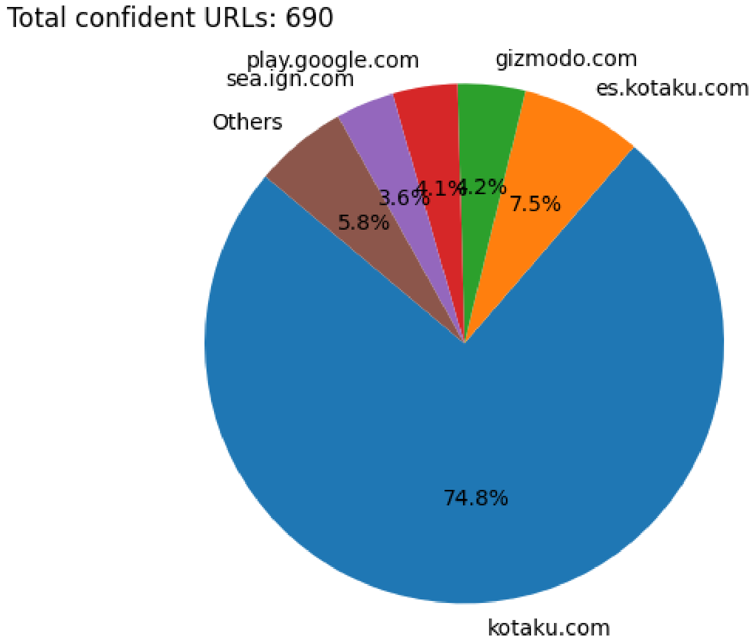


Figure 5: Breakdown (by percentage and domain) of URLs visited related to gaming