Scraping the Internet’s Most Popular Websites

**I am not a lawyer. This blog post is not legal advice.**

Last week, I collected a list of the most popular websites from [Wikipedia](https://en.wikipedia.org/wiki/List_of_most_popular_websites) and scraped them. Granted I only scraped where all robots should be welcome, robots.txt.

For those unfamiliar, various companies/individuals use automated programs (robots) to collect data and robots.txt is a file that many websites have that regulate where robots are allowed on their site. You can find a site’s robots.txt by entering in the url + /robots.txt. For example, here is Google’s robots.txt.



User-agent designates which bot the directions belong to (User-agent: \* means all users). Disallow tells robots where not to go, and allow tells robots where they may go.

I decided to scrape these websites to find which ones may be open to scraping based on the rules of their robots.txt. Now if you’re considering webscraping, you should check the website’s terms of service and search whether the company your scraping from is litigious on these kinds of matter. Better yet – talk to a lawyer!

However, I would think that a website’s robots.txt would conform to their policy on webscraping because ultimately robots that are crawling multiple sites probably are not analyzing the terms of service, but if built properly they are definitely reading and obeying the robots.txt file.

The analysis below does not cover specific user agents, and only pertains to User-Agent: \* (rules for all users). I deduplicated the list of websites from [Wikipedia](https://en.wikipedia.org/wiki/List_of_most_popular_websites) when multiple variants of the same site were included (I.e. there was Google, Google Canada, Google Hong Kong, etc.). The final list was 84 sites and is [linked here](https://github.com/brendanbailey/Medium/blob/master/robots_txt/wikipedia_popular_sites.csv). Note that this data was collected on April 19th, and a website’s robots.txt may have changed from then.



Most sites were classified as Mixed because they have a mixture of rules of where robots can go. Google’s robots.txt is an example of a mixed robots.txt.

Complete Allow means that the site’s robots.txt file explicitly gives robots complete access (i.e. Disallow: ). Probably the most notable site from my perspective is SoundCloud.



**Here is the full list of Complete Allow Sites:**

360.cn

fc2.com

livejasmin.com

popads.net

qq.com

soundcloud.com

uol.com.br

Not set means either the site’s robots.txt does not have rules for User Agent: \* or that the site does not have a robots.txt file. Whether intended or not, this is the same as granting robots complete access to their site. The most notable site on this list is Diply.



**Here is the full list of sites where robots.txt is not set:**

cnzz.com

diply.com

jd.com

live.com

pixnet.net

sina.com.cn

sohu.com

tianya.cn

vk.com

Complete disallow sites explicitly prohibit robots from accessing their site (i.e. Disallow: /). This is a pretty clear signal to avoid scraping these sites. A notable site that falls into this category is LinkedIn, [who has engaged in lawsuits with webscrapers](https://betanews.com/2016/08/16/linkedin-sues-data-scrapers/).

**Here is the full list of Complete Disallow Sites:**

facebook.com

go2cloud.org

hao123.com

linkedin.com

Naver.com

onclckds.com

soso.com

t.co

taobao.com

tmall.com

Hopefully the above is helpful in looking at where to start and avoid for your various webscraping projects. If you would like to how I collected this data or see the results for all the sites I collected data from, [you can check my repo](https://github.com/brendanbailey/Medium/tree/master/robots_txt). As stated above, robots.txt is not the only factor in deciding to scrape, and in case it wasn’t clear: **I am not a lawyer. This blog post is not legal advice.**