Get a list of robot user agents from

Web crawlers Json file on Github

filter the json on user agent key

```
jq -c 'map(with_entries(select(.key == ("pattern"))))'
```

==> robot_patterns.json

get the user agent from each line

```
grep ^(?!"pattern").*$
```

==> robot_patterns.txt

filter to leave only bot names present in data (fish shell)

Input `data_user_agents.txt` is simply the use agent column of the given `clickdata.csv`

- loop over each line
- print occurrences of robot patterns
- count unique occurrences of robot patterns

- sort by count

```
for line in (cat robot_patterns.txt); grep -oh $line data_user_agents.txt ; end | uniq -cd | sort -nr > ./data_uniq
```

To Tab-delimited csv

Replace space by tab

```
cat ./data_uniq | sed 's/^\s*//' | sed 's/ /\t/g;s/ /,/' > data_uniq.csv
```

Manually label the bot names as NHT-Search or NHT-Other

==> labeled_web_crawlers.csv

Count	Name	Type	Info
18123	Googlebot	NHT-search	
1032	AdsBot-Google-Mobile	NHT-other	
670	Mediapartners-Google	NHT-other	
234	Sogou	NHT-search	

208	bingbot	NHT-search	
171	Applebot	NHT-other	
65	facebookexternalhit	NHT-other	
48	WhatsApp	NHT-other	
34	UptimeRobot	NHT-other	
33	pinterest	NHT-other	
32	YandexBot	NHT-search	
17	SemrushBot	NHT-other	
11	Googlebot-Image	NHT-search	
10	Gluten Free Crawler	NHT-other	Joke: searches for websites containing names of food products containing gluten
4	coccoc	NHT-other	Vietnamese browser
3	Twitterbot	NHT-other	
3	Facebot	NHT-other	
3	Discordbot	NHT-	

		other	
3	Cliqzbot	NHT-other	
2	Google Web Preview	NHT-other	
2	AppEngine-Google	NHT-other	