

SEO - Používáme robots.txt

Soubor robots.txt se umísťuje zvyčajne do koreňovej zložky stránky. Pomocou neho dokážete zakázať robotom prístup k určitým súborom alebo adresárom, ktoré nechcete aby vyhľadávač zaindigoval. Niekedy napríklad nejde ani o robotov ale skôr o úšetrenie prenesených dát vášho hostingu. Taký roboti prenesú vysoké číslo prenesených dát. Len si predstavte koľko to musí byť, pokiaľ navštíví každú URL na stránke.

Robots.txt musí mať ukončovanie riadkov typu UNIX (LF). Takže nezabudnite súbor uložiť v správnom formáte! Myslím, že nemusíte byť programátorom na to, aby ste si mohli dovoliť pracovať s týmto obyčajným textovým súborom. Ako prvé čo sa do robots.txt píše je text User-agent:

Určíme ním názov robota, napr.:

```
User-agent: googlebot
```

Samozrejme robotov máme na svete obrovské množstvo ale tých dôležitých je len niekoľko s nich. Ak chcete nastaviť robots.txt pre všetkých robotov, stačí ak zapíšete do User-agent: nasledujúci parameter:

```
User-agent: *
```

Pod názov robota napíšeme na nový riadok Disallow:

Tento zápis slúži na zakázanie/povolenie prístupu robotov na celé stránky, či len určité zložky a súbory.

```
User-agent: *  
Disallow:
```

Tento príklad špecifikuje zápis súboru, tak aby všetci roboti navštívili každý súbor alebo zložku na stránke.

Naopak nasledujúci zápis zakáže všetkým robotom prístup.

```
User-agent: *  
Disallow: /
```

Zakázať robotom snoriť v zložkách môžete takto:

```
User-agent: *  
Disallow: /zlozka1/  
Disallow: /zlozka2/  
Disallow: /zlozka3/
```

Žiadny robot teraz nebude navštevovať zložky zložka1, zložka2, zložka3.
V robots.txt máme ešte ďalšiu možnosť zápisu a je to Crawl-Delay.
Viackrát Crawl-Delay: môžeme nastaviť v sekundách dobu, ktorú robot strávi na stránke.

Na začiatku som písal ako možno ušetriť prenesené dáta. Teraz sa k tomu vrátim a napíšem riešenie, ktoré zabráni robotom čerpať veľké množstvo dát.
Tento screenshot pochádza z mojích štatistík návštevnosti. Všimnite si koľko trafficu čerpajú "len" roboti.

Všimnite si robota, ktorý čerpá najväčšie množstvo dát. Stránka sa mu páči natoľko, že na nej trávi neprimerane veľa času :-). V skutočnosti je to asi tým, že navštevuje stránku príliš často. Aby sme zabránili takémuto "vyžieraniu" dát, mali by sme robota obmedziť maximálnou prístupovou dobou.
Preto použijeme Crawl-delay.

```
User-agent: Slurp  
Crawl-delay: 15
```

Podľa zápisu si už teraz robot Slurp na stránke pobdnie maximálne 15 sekúnd, čo predpokladám je dosť na to, aby nečerpal toľko dát koľko nemusí. Ak zapojíme trochu logiky, môžeme takéto opatrenie vykonať pre všetkých robotov takto:

```
User-agent: *  
Crawl-delay: 15
```

Toto sa však neodporúča a preto by bolo lepšie keby ste to nepoužívali. Ak si chcete skontrolovať váš robots.txt validátorom, tak máte na výber validátor na stránke SearchEngineWorld, alebo vám pomôže Google.
Nenechajte sa hneď zmiasť, ak validátor vyhodí chybu napr. pri Crawl-delay.
Crawl-delay validné nie je, avšak funkčné určite je (záleží od robota). Je to len mimo štandard. Taký Googlebot Crawl-delay ignoruje ale napríklad Slurp či msnbot ho podporujú.

Ako každý správny skript (toto síce nie je skript), robots.txt ponúka možnosť zapísania komentárov. Komentáre sa zapisujú rovnako ako napr. pri úprave httpd.conf v Apache, čiže takto:

```
#Roboti sú na tejto stránke vítaní  
User-agent: *  
Disallow:
```

Ako som sa na internete doèítal urèite, nezakomentuje niè takýmto spôsobom:

User-agent: *
Disallow: nieco #tu mam subory

Väèšina robotov bude pracova• s komentármi v robots.txt správne ale dôvod preò by ste nemali komentova• riadky takýmto spôsobom je ten, že niektorí roboti budú stránku indexova• takto:

User-agent: *
Disallow: nieco#tu mam subory

Dajte si na toto pozor! Nikdy neviete aký robot vás navštívi. Venujte pozornos• aj názvu súboru, pretože správne môže by• len robots.txt, a nie nejaké ROBOTS.txt, èi ROBOTS.TXT. Proste case-sensitive.

Odkazy s ěalšími zdrojmi informácií:

- <http://www.webmasterworld.com/forum93/>
- http://www.searchengineworld.com/robots/robots_tutorial.htm
- <http://www.searchengineworld.com/cgi-bin/robotcheck.cgi>
- <http://www.robotstxt.org/wc/robots.html>
- Google[Robots.txt]