robots.txt چیست ؟ چگونه آن را ایجاد و مدیریت کنیم ؟

نام فایل : آموزش robots.txt

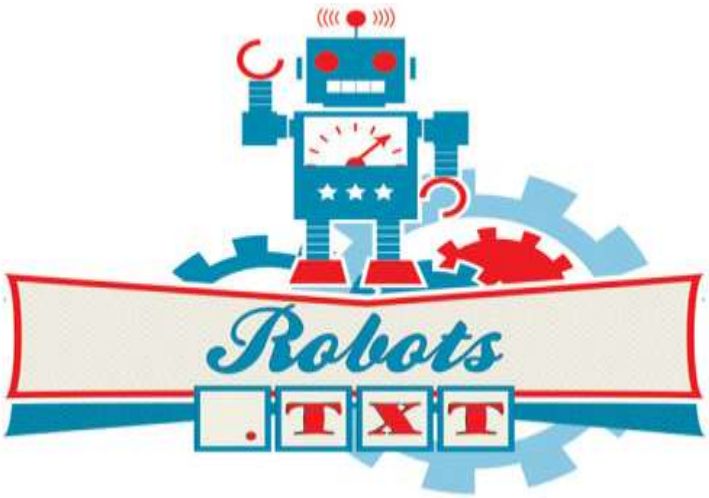
دسته آموزشی : سئو و ایندکس

نویسنده آموزش : مهرداد رستمی ینگجه

تاریخ نگارش مقاله : 28 بهمن 1394

تعداد صفحات : 8

وب سایت : *GhalebGraph.ir*



Robots.txt

فایل robots.txt مهم ترین وسیله در سئو به منظور افزایش سرعت ایندکس مطالب سایت شما در موتورهای جستجو می باشد . به وسیله این فایل میتوانید تعیین کنید ربات های خزنده موتورهای جستجو چه محتوایی از سایت شما را ایندکس نکنند برای ایجاد فایل ربات ابتدا نوت پد ویندوز خود را باز کنید ، سپس دستوراتی به مانند زیر را وارد کنید :

```
User-agent: *
Disallow: /cgi-bin/
Allow: /cgi-bin/cat/
Crawl-delay: 10
Sitemap: http://www.yoursite.com/sitemap.xml
```

خب هم اکنون دستورات به کاربرده شده در نمونه بالا را تک به تک و به ترتیب برای شما یاوران قالب گراف توضیح میدهیم.

User-agent

این تابع مشخص کننده نوع رباتی است که دستوراتی برای آن جهت اجرا تعیین می شود . می توان جلوی این تابع علامت * یا سایر مقادیر مخصوص که مربوط به موتور جستجوی خاصی است به کار برد که در نمونه بالا از * استفاده کردیم.

نکته : * خط فرمان به تمام ربات ها برای تعیین دستورات در فایل ربات می باشد.

جدول نام مقادیر مخصوص برای موتور های جستجو در فایل ربات

نام موتور جستجوگر	نام متغیر در ربات	نام موتور جستجوگر	نام متغیر در ربات
گوگل	googlebot	Cuil	twiceler
یاهو	yahoo-slurp	Scrub The Web	scrubby
MSN Search	msnbot	Nutch	nutch
Ask/Teoma	teoma	Baidu	baiduspider
DMOZ Checker	robozilla	GigaBlast	gigabot
Alexa/Wayback	ia_archiver	Naver	naverbot, yeti

نکته : البته تعداد بیش از 300 موتور جستجو در حال حاضر از فایل ربات پشتیبانی می کنند؛ مانند WebReaper و Offline Explorer و IsraBot (دائره المعارف وی کی پدیا) و ... که در جدول بالا به آنها اشاره نکرده ایم. برای مشاهده لیست کامل متغیر های مربوط به جستجوگر ها به آدرس زیر مراجعه نمایید :

http://www.robotstxt.org/db.html

جدول نام مقادیر مخصوص برای قسمت های خاص برخی موتور های جستجو در ربات

نام موتور جستجوگر	نام متغیر در ربات	نام موتور جستجوگر	نام متغیر در ربات
Google Image	googlebot-image	Yahoo MM	yahoo-mmcrawler
Google Mobile	googlebot-mobile	Yahoo Blogs	yahoo-blogs/v3.9
MSN PicSearch	psbot	SingingFish	asterias

به عنوان مثال برای دو موتور جستجوی مختلف به این شکل رباتی تنظیم کرده ایم :

```
User-agent: yahoo-slurp
Disallow: /category/
User-agent: Googlebot
Disallow: /oldposts/
```

معنای مثال بالا این است که موتور جستجوی یاهو به پوشه category و موتور جستجوی گوگل هم به پوشه oldposts دسترسی نداشته باشد.

نکته : پس از تعیین نوع User-agent هر دستوری زیر آن بیایید متعلق به آن است و اگر بار دوم User-agent در ربات بیاید دستورات تایپ شده بعد از User-agent دوم فقط متعلق به همان دومی است نه اولی. به صورت کلی هر دستور تایپی پایین User-agent مربوط به User-agent بالایی آن می باشد و به دیگر User-agent های موجود در ربات تعلق ندارد.

Disallow

این تابع نشان دهنده آدرس صفحه ای است که میخواهید از دید روبات ها پنهان بماند. توجه نمایید در شروع آدرس ها از کاراکتر اسلش/ استفاده کنید و همچنین دقت کنید که لازم است در پایان نام فولدر ها از کاراکتر / استفاده شود. هر آدرسی در هر سطر مقابل Disallow نوشته شود به این معناست که ادرس سایت شما + آن آدرس ایندکس نشود!

```
User-agent: *
Disallow: /picture/
Disallow: /about.html
Disallow: /images/mypic.jpg
```

در مثال کادر بالا ما تعیین کردیم که اگر ادرس سایت شما به این صورت باشد yorname.ir در این صورت محتویات فولدر (پوشه) یا مسیر yorname.ir/picture/ و صفحه yorname.ir/about.html و عکس yorname.ir/images/mypic.jpg توسط همه موتور های جستجو بررسی نگردد و ایندکس نشود.با مثال های قالب گراف میتوانید راحت بیاموزید !

نکته : از دستور زیر برای مسدود کردن دسترسی تمام موتور های جستجو به تمام صفحات سایت استفاده میگردد.

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

نکته : از دستور زیر برای دسترسی تمام موتورهای جستجو به تمام محتویات سایت استفاده میگردد.

```
User-agent: *
Allow: /
```

نکته : از دستور زیر برای عدم دسترسی جستجوگر خاص به تمام محتویات سایت استفاده میگردد. (به جای Badbot نام متغیر مربوطه به جستجوگر مد نظر خود را در مثال زیر وارد نمایید به عنوان مثال googlebot و ...)

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

Allow

این دستور کم تر در فایل ربات به کار گرفته میشود . فلسفه استفاده از این دستور به این مسئله برمی گردد که وقتی شما یک مسیر را مسدود میکنید تا ربات های موتور جستجو به آن دست نیابند اما چندین مسیر درون آن مسیر مسدود شده وجود دارد که نیاز به ایندکس آنها می دانید از این دستور استفاده میکنید. به عنوان نمونه ، مثال زیر را داریم :

```
User-agent: *
Disallow: /picture/
Allow: /picture/boy/
```

مفهوم مثالی که در بالا سایت قالب گراف برای شما آورده است به این معناست که مسیر picture/ و تمام مسیر های درون این مسیر از سایت شما توسط هیچ موتور جستجویی ایندکس نشود. اما از داخل هزارن مسیری که احتمال دارد درون مسیری که سابقا آن را مسدود کردیم وجود داشته باشد ، مسیر picture/boy/ ایندکس شود و از بند مسدود سازی خارج گردد. چنین دستوری هنگامی کمک حال ما میشود که بخواهیم از داخل یک پوشه که هزاران پوشه دیگر دارد فقط چندین مسیر خاص را مسدود نماییم که امکان تک به تک مسدود سازی آن هزار پوشه و فقط مسدود نکردن چندین پوشه خاص به دلیل دستورات زیادی که خواهیم نوشت ، سخت است ، که در این شرایط از دستور فوق برای کوتاه کردن دستورات و آسانی کار استفاده میکنیم.

Crawl-delay

این دستور ، یک دستور تاخیری است که در اغلب جستجوگر ها به جز برخی همانند گوگل و ... به کار برده میشود. اغلب افراد کم تر از این دستور در ربات خود استفاده میکنند چون فقط گوگل برای آنها اهمیت دارد . این تابع ، سرعتی که هر ربات می تواند یک سرور را کرال کند به میلی ثانیه نشان می دهد. در گوگل باید برای تنظیم این مورد وارد وبستر خود شوید.

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

Sitemap

پس از نوشتن تمام دستورات ، در آخرین سطور فایل ربات این دستور را استفاده میکنیم و مقابل این دستور آدرس نقشه سایت خود یا همان سایت مپ را وارد کنیم تا با این روش نقشه سایت خود را موتور های جستجو معرفی نماییم .

```
Sitemap: http://YourDomain.com/sitemap.xml
```

یادآوری این موضوع نیز حائز اهمیت است که شما اگر بیش از یک سایت مپ داشته باشید میتوانید همه آنها را سطر به سطر معرفی کنید . بیشتر افراد فقط در ربات ادرس سایت مپ اصلی خود را وارد میکنند.

نحوه ذخیره سازی و ایجاد فایل ربات در سیستم های مدیریت محتوا

در نهایت پس از تصمیم گیری برای این که چه مسیر هایی را باید غیر قابل دسترس کنید و نوشتن این دستورات در فایل نوت پد این فایل را با نام robots با فرمت txt ذخیره میکنیم. سپس این فایل را در روت هاست خود آپلود میکنیم. تا ادرس ان به صورت زیر قابل لود باشد:

http://www.example.com/robots.txt

به یاد داشته باشیم ادرس دهی به این شکل بسیار مهم است.

نحوه ذخیره سازی و ایجاد فایل ربات در سرویس های وبلاگدهی

در سرویس های وبلاگدهی این فایل اکثرا توسط خود مالکان سرویسهای وبلاگدهی برای هر سایت ساخته میشود و امکان ویرایش آن نیست. لازم به ذکر است که در این گونه سیستم ها محتویات فایل ربات برای همه وبلاگ های آن سیستم یکسان است. اما در این بین تا تاریخ نگارش این مطلب فقط در سرویس وبلاگدهی رزبلاگ امکان ویرایش این فایل توسط خود کاربران وجود دارد.

در سرویس وبلاگدهی رزبلاگ جهت ویرایش فایل ربات خود در پنل سایت خود در قسمت ((تنظیمات)) روی ((تنظیمات robots.txt)) کلیک کنید و از آن قسمت فایل ربات سایت خود را طبق آموزشی که دادیم ویرایش و ثبت کنید.

ساختی رباتی پیشرفته تر

برخی از جستجوگر های پرقدرت همانند گوگل از چندین قاعده دیگر نیز در ربات پیروی میکنند که به وسیله آنها میتوان ربات پیشرفته تری ایجاد کرد. با استفاده از سه علامت (* ? $) میتوانیم این کار را انجام دهیم.

* برای نشان دادن تطبیق توالی در نام استفاده میشود.

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

در مثال بالا منظور آن است که هر پوشه ای که با کلمه boy شروع شود مانند boyclothes و boylearn و ...توسط جستجوگر گوگل ایندکس نشود.

*? برای مسدود سازی دسترسی به صفحات دینامیک استفاده میشود.

```
User-agent: msnbot
Disallow: /*?
User-agent: Googlebot
Disallow: /books/*?
```

در مثال بالا دسترسی جستجوگر ام اس ان به کل صفحات دینامیک و دسترسی جستجوگر گوگل به صفحات دینامیک پس از این آدرس /books/ فقط مسدود شده است.

$ برای مشخص کردن انتهای یک الگو استفاده میگردد.

```
User-agent: Googlebot
Disallow: /*.gif$
```

بر طبق مثال بالا دسترسی جستجوگر گوگل به تمام تصاویری که با فرمت gif در سایت می باشند مسدود شده است.

چرا ایجاد فایل ربات مهم است ؟

1. باعث افزایش سرعت ایندکس و بررسی صفحات سایت شما توسط موتور های جستجو می شود.

2. اکثر موتور های جستجو همواره این فایل را از سرور شما درخواست میکنند بنابراین برای جلوگیری از بروز خطای 404 در موتور های جستجو برای سایت شما ، ایجاد آن بهتر است.

نکاتی مهم در خصوص فایل ربات

1. فایل ربات باید در ریشه سایت با نام robots.txt ایجاد شود.

http://www.example.com/robots.txt

2. هر دستور را در یک سطر باید بنویسم و نوشتن پشت سر هم دستورات باعث عدم اجرای دستورات می شود. (مثال زیر نمونه اشتباه است)

Disallow: /cgi-bin/ /tmp/

3. برخی از ربات ها، مخصوصا ربات های مخرب (Malware) می توانند فایل robots.txt شما را نادیده بگیرند و اصلا به آن توجهی نکنند . برخی از ربات های مخرب از اینکار استفاده کرده اند تا قادر باشند وب سرور را برای مشاهده ضعف های امنیتی اسکن نمایند و یا آدرس ایمیل های مفید را برای افراد اسپمر جمع آوری کنند .

4. در استفاده از حروف بزرگ و کوچک دقت کنید چرا که جستجوگر ها به این مورد حساس هستند به این معنا که مسیر file با File متفاوت است.

5. برای حفظ حریم خصوصی و اطلاعات خانوادگی از این فایل استفاده نکنید چرا که عموم قابلیت دسترسی به این فایل را دارند و میتوانند سو استفاده کنند. پس برای حفاظت از عکس های خانوادگی و ... در سایتتان ، از فایل ربات استفاده نکنید.

مسدود سازی به روشی دیگر

در بالا خاطر نشان کردیم که هر صفحه ای را به علت سودجویی افراد یا موتورهای جستجو نباید در ربات سایت قرار داد . برای مسدود سازی چنین صفحاتی که در ربات بنا به دلایل ذکر شده نتوانستیم آن را مسدود (از دسترسی موتور های جستجو محافظت) کنیم می توانیم در قالب آن صفحه از کدی به مانند زیر استفاده کنیم که باعث عدم دسترسی موتورهای جستجو به محتویات آن صفحه می شود.

```
<meta name="robots" content="noindex" />
یا
<meta name="googlebot" content="noindex" />
```

در کد بالا، کد موجود در سطر اول کاربردش این است که دسترسی همه موتور جستجو به آن صفحه مسدود شود . اما در کد دوم متغیر گوگل تعیین شده است که فقط دسترسی جستجوگر گوگل را به صفحه محدود می کند. در کد دوم به جای googlebot می توانیم نام مخصوص سایر موتور های جستجو را که قبلا در آموزشمان معرفی کردیم ، قرار دهیم.

نمونه های آماده ربات برای سیستم های مدیریت محتوا و سرویس وبلاگدهی رزبلاگ

در این قسمت از آموزش ،سایت قالب گراف، برای سیستم های مدیریت محتوا و سرویس وبلاگدهی رزبلاگ فایل ربات پیشنهادی ارائه میدهد. در صورت تمایل میتوانید از این دستورات استفاده نمایید.

Wordperss

*User-agent: *
Disallow: /wp-admin
Disallow: /wp-includes
Disallow: /wp-content/plugins
Disallow: /wp-content/cache
Disallow: /wp-content/themes
Disallow: /trackback
Disallow: /tag
Disallow: /author
Disallow: /wget/
Disallow: /httpd/
Disallow: /cgi-bin
Disallow: /images/
Disallow: /search
Disallow: /feed
Disallow: /feed/
Disallow: /trackback/
Disallow: /rss
Disallow: /comments/feed
Disallow: /feed/$
Disallow: /*/feed/$
Disallow: /*/feed/rss/$
Disallow: /*/trackback/$*

joomla

*User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /editor/
Disallow: /help/
Disallow: /includes/
Disallow: /language/
Disallow: /mambots/
Disallow: /media/
Disallow: /modules/
Disallow: /templates/
Disallow: /installation/
Disallow: /libraries/
Disallow: /tmp/
Disallow: /xmlrpc/
Disallow: /admin
Disallow: /administrator
Disallow:/admin/
Disallow: /admin.html
Disallow:/admin.php*

Rozblog

*User-agent: *
Disallow: /vote
Disallow: /User/
Disallow: /user/
Disallow: /Polls/
Disallow: /rating/
Disallow: /images/smilies/
Disallow: /Quote/
Disallow: /include/captcha/
Disallow: /Forum/User/
Disallow: /Forum/Add/
Disallow: /Forum/Tanks/
Disallow: /Forum/Thanks/
Disallow: /Forum/New_Post/
Disallow: /Forum/New/Answer/
Sitemap: http://Yoursite.ir/sitemap.xml*