

Reflektionsfrågor

Finns det några etiska aspekter vid webbskrapning. Kan du hitta något rättsfall?

1. Webbskrapning kan orsaka försämrade prestanda på webbsidor. Detta kan med flit nyttjas för att orsaka så kallade "Denial of Service"-attacker. Dessa attacker utförs exempelvis vid utpressning, hämnd eller elektronisk krigföring.
2. Webbskrapning kan nyttjas till att skrapa någon annans verk på en webbsida för att sedan posta det som sitt eget. Detta är ren stöld. Läs alltid på vilka rättigheter du har när du använder någon annan persons verk.

Rättsfall

1. eBay v. Bidder's Edge: Fallet handlade om att bud lades per automatic i sista stund (så kallat "Auction Sniping") på eBays webbsida. Fallet resulterade i ett domstolsbeslut att Bidder's Edge fick sluta hämta, samla och indexera aktioner från eBay's webbsida.

Finns det några riktlinjer för utvecklare att tänka på om man vill vara "en god skrapare" mot serverägarna?

1. Läs igenom webbplatsens "Terms of use".
2. Se till att följa reglerna i robots.txt-filen.
3. Lämna alltid någon typ av signatur när du skrapar. Uppge exempelvis vem du är samt ett par kontaktuppgifter.
4. Låt det alltid ta några sekunder mellan alla dina requests för att inte råka överbelasta serverna.

Begränsningar i din lösning- vad är generellt och vad är inte generellt i din kod?

Generellt

1. Ej beroende av antalet personer.
2. Ej beroende av antalet dagar vid skrapning. Dock checkar jag enbart mot helgdagarnas namn när lediga dagar sparas.
3. Ej beroende av antalet undersidor till huvudsidan.
4. URL:erna är delvis generella. Vid GET av film och dag på filmsidan är en del av URL:en hårdkodad. "/check?day=", "&movie=".

Inte generellt

1. Webbskrapan är beroende av att ha hittat undersidorna vid namn "calendar, cinema, dinner" för att kunna gå vidare i koden.
2. Webbskrapan är beroende av att inga ID:n ändras.

3. Webbskrapan är beroende av att de olika taggarnas "values". Exempelvis input-taggen.
4. Webbskrapan är beroende av de värden som sats i JSON-strängen vid verifiering av lediga filmer.
5. Webbskrapan är beroende av att ordet "OK" fyllts i kalendern för att räkna detta som en tillgänglig dag.

Vad kan robots.txt spela för roll?

Det är en textfil i rootkatalogen som talar om vad du får och inte får skrapa från webbplatsen. Den hjälper dig att inte råka bryta mot lagen/eventuella regler.