NAVN

webanalyzer − tekst analyser et websted

SYNOPSIS

webanalyzer *url* [**-d** *depth*] [**-u** *agent*] [**-c** *delay*] [**-lix**] [**-fre**] [**-fkgl**] [**-wordrep**] [**-l** *language*] [**-o** *directory*] [**-ignore-tag** *html_tag*] [**-ignote-id** *id*]

INSTALLATION

webanalyzer er afhængig af tre filer for at køre - det primære program, en fil ved navn **webanalyzer,** og to hjælpefiler, **smlnjruntime** og **webanalyzer.x86-linux,** og alle disse tre filer skal ligge i samme mappe for at programmet fungerer. Kopiér filerne ind i et praktisk sted og kør **webanalyzer** for at afvikle programmet.

BESKRIVELSE

webanalyzer er et program til at analysere sværhedsgraden af tekst på hjemmesider. Dette gøres ved brug af forskellige tekst analyser og stavekontrol:

LIX er en forkortelse for læsbarhedsindeks og er en skala for en given teksts læsbarhed. Dette opgøres som det gennemsnitlige antal ord pr. helsætning, plus procentdelen af lange ord, altså ord der er over seks bogstaver lange. LIX blev introduceret af den svenske pædagog C.H. Björnsson (1916-1988). Jo større lixtallet altså er, desto sværere regnes teksten for at være.

Følgende skala bruges til at vurdere uddata:

>55	Meget svær, faglitteratur på akademisk niveau, lovtekster.
45-54	Svær, f.eks. saglige bøger, populærvidenskabelige værker, akademiske udgivelser.
35-44	Middel, f.eks. dagblade og tidsskrifter.
25-34	Let for øvede læsere, f.eks. ugebladslitteratur og skønlitteratur for voksne.
<24	Let tekst for alle læsere, f.eks. børnelitteratur.

Kilde: http://da.wikipedia.org/wiki/LIX

FKGL bruges bl.a. inden for uddannelse, som er et af de mest åbenlyse steder at bruge læsbarheds analyser. FKGL (Flesch-Kincaid Grade Level) overfører resultatet (inden for 0-100) til det Amerikanske uddannelses trin, hvilket gør det nemmere for lærer, forældre, bibliotekarer og andre at bedømme uddannelses niveauet for forskellige bøger og tekster. Det kan også tolkes som antal års generel uddannelse som er krævet for at forstå teksten (Dette er brugbart når teksten giver resultat over 12.

Kilde (Frit oversat):
http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

FRE som er en forkortelse for Flesch Reading Ease. Højere resultater indikerer materiale som er let læseligt og lavere resultater markerer sværere materiale. Som en tommelfinger regel, er tekster på 90.0−100.0 tænkt let forstået af en gennemsnits 5. klasse. 8. og 9. klasses studerende kan nemt forstå passager med resultat på 60.0−70.0 og passager på 0.0−30.0 er bedst forstået af universitets studerende. Fx. har Reader's Digest magazine et læsbarheds indeks på ca. 65, Time magazine har omkring 52 og Havards Law Review har en generel læsbarheds indeks på 30. Denne test er blevet en Amerikansk regerings standard. Mange regerings afdelinger kræver at dokumenter eller former har et vis læsbarheds indeks bl.a. "U.S. Department of Defense". De fleste stater kræver at forsikrings former skal have et resultat på 40.0−50.0. Brug af denne test er så udbredt at den ofte findes i populære skrive programmer som KWord, Lotus WordPro og Microsoft Word. Lange ord påvirker dette analyse væsentlig mere end de gør i **FKGL**.

Kilde (Frit oversat):
http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

Sidesværhedsgrad er en sammensat sværhedsgrad for siden i en helhed. Når analysen er færdig og der er genereret en indeks fil vil alle de analyserede sider stå i en tabel sorteret efter deres sidesværhedsgrad. Sidesværhedsgraden markeres ydermere med en baggrunds farve som går fra lys grøn (let) hen i gul (middel) og slutter i mørk rød (svær). Sidesværhedsgraden for de individuelle sider i indeks filen er hele sidens samlede værdi. Når hver enkelt analyserede fil åbnes for sig vil den analyserede teksts individuelle afsnit blive vist med en farvet kasse rundt om som afspejler sidesværhedsgraden for afsnittet selv. Teksten i hvert enkelt afsnit er også farvet med en baggrundsfarve som afspejler sidesværhedsgraden for den enkelte sætning.

Stavekontrol bliver udført på alle analyserede ord. teksten analyseres tjekkes der for attributten "lang" i alle HTML tags. Hvis et HTML tag indeholder denne attribut overskrives standard sproget for al tekst indkapslet af dette tag. Da stavekontrollen bruger `aspell`(1) er det krævet at denne er installeret samt ordbøger til de sprog som forventes analyseret. Hvis der ikke forefindes nogen ordbog for det givne sprog der analyseres vil der ikke blive reporteret nogle fejl, blot vil alle ord se ud

som om de var stavet korrekt (Altså ingen ændringen i dokumentet grundet stavekontrollen). Når stavekontrollen finder et ord som ikke er stavet korrekt sættes en blå kasse rundt om ordet i uddata. It Gentagne ord bliver altid udført på den analyserede tekst. Analysen markerer det gentagne ord med en orange kasse i uddata.

Alle tekstanalysernes brugbarhed afspejles i den analyseres teksts størrelse. For at brugbarheden er optimal skal den analyserede tekst være omkring 1.000 ord eller mere. Herved skal sidesværhedsgraden for enkelte sætninger i afsnit ikke vægtes højt men blot ses som en indikation af at en enkelt sætning eller et afsnit kan være svært.

OUTPUT

Når programmet har analyseret et websted gemmes alle filerne i en mappe, enten specificeret ud fra "**-o** *dir*" parametren (se nedenfor) eller også laves en mappe (i det bibliotek som programmet blev startet fra) som navngives efter det domæne som er blevet analyseret. Når en enkelt fil er blevet analyseret bliver den gemt i mappen og efter endt analyse bliver der skrevet en indeks fil kaldet "index.html" hvor der er links til alle analyserede filer. Listen i indeks filen er sorteret efter højest sidesværhedsgrad i toppen ydermere er alle elementerne er også farvet ud sin sidesværhedsgrad. Når en enkelt side er analyseret og skal gemmes bliver alle "/" erstattet med "#" og alle filer bliver suffikset med ".html" så fx filen: "http://dybber.dk/test/test.html" bliver gemt i mappen som: "dybber.dk#test#test.html.html". Grunden til dette suffiks er at hvis den analyserede fil endte på ".php" eller ".jsp" vil de fleste browsere ikke åbne dem da de logisk nok ikke tror det er html filer.

Når "index.html" åbnes og der klikkes videre til en konkret side vil man se den enkelte sides analyse resultat. Resultatet er en strippet udgave af den originale side hvor kun tekst (der ikke er indkapslet af tags som er bedt ignoreret) optræder tilbage. Den oprindelige tekst er her blevet delt i afsnit som hver for sig vist sine analyse resultater. Hvert afsnit har herved en farvet kasse rundt om, som indikerer sidesværhedsgraden for afsnittet. I toppen af kassen på det individuelle afsnit står resultaterne for de individuelle tekst analyser som er valgt. Under analyse resultaterne for det individuelle afsnit er teksten som er har farvet baggrund alt de individuelle sætningers sidesværhedsgrad. De individuelle ord kan ydermere også have specielle farver. Hvis fx der er en blå kasse rundt om indikerer dette at ordet ikke er stavet korrekt og hvis ordet har pink baggrund er det fordi det er et gentagende ord.

KOMMANDOLINJEPARAMETRE

url Angivelse af den internetadresse (URL) som **webanalyzer** skal analysere. Adressen skal præfikses med "http://" og kan enten angive et websted fx. "http://google.dk" eller en specifik side fx. "http://sigkill.dk/index.html".

-d *num*
Sætter dybden af links som crawleren skal følge fra startsiden. En dybde på 0 vil sige at kun forsiden analyseres.

-u *str*
Sætter den User-agent som sendes med alle HTTP requests og som der identifiseres med i robots.txt filer

-c *num*
Sætter pausen mellem hver HTTP request.

-lix Skifter hvorvidt der skal bruges lix analyse på teksten. Læs mere: http://da.wikipedia.org/wiki/LIX

-fre Skifter hvorvidt der skal bruges 'Flesch Reading Ease' analyse på teksten. Læs mere: http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

-fkgl
Skifter hvorvidt der skal bruges "Flesch-Kincaid Grade Level" analyse på teksten. Læs mere: http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

-wordrep
Skifter hvorvidt der skal tjekkes for gentagne ord.

-l *lang*
Sætter default sprog kode som bruges hvis der ikke er angivet noget andet i det analyserede dokument.

-o *dir*
Sætter output mappen til en relativ eller absolut sti. Hvis der ikke er angivet nogen sti bliver der oprettet en mappe (navngivet efter det domæne der er angivet som URL) i den mappe som programmet startes fra.

-ignore-tag *tag*
Angiver hvilket html tag der skal filtreres fra i analysen. Kan defineres flere gang hvis flere tags ønskes filtreret fra.

-ignore-id *id*

Angiver hvilket html tag med givet id der skal filtreres fra i analysen. Kan defineres flere gang hvis flere tags ønskes filtreret fra.

EKSEMPLER

Analyser et websted med **FRE** og **FKGL** analyserne skiftet:

```
$ webanalyzer http://host.dk -fre -fkgl
```

Analyser et websted og brug "Tekst_Analyse" som user-agent i stedet for standard:

```
$ webanalyzer http://host.dk -u Tekst_Analyse
```

Analyser et websted og begræns crawlingen med dybde 3 og vent 10 sekunder mellem hver http request:

```
$ webanalyzer http://host.dk -d 3 -c 10
```

Analyser et websted og gem uddata i et andet bibliotek og tjek for gentagne ord:

```
$ webanalyzer http://host.dk -o /home/bruger/valgt_mappe -wordrep
```

Analyser et websted og ignorer al data der ligger i HTML <table> tags

```
$ webanalyzer http://host.dk -ignore-tag table
```

Analyser et websted og brug engelsk stavekontrol som standard og ignorer al data som ligger i HTML tags med attribut id som har værdien “menu”

```
$ webanalyzer http://host.dk -l en -ignore-id menu
```

FORFATTERE

Troels Henriksen ⟨athas@sigkill.dk⟩ Martin Dybdal ⟨dybber@dybber.dk⟩ Jesper Reenberg ⟨reenberg@kampsax.dtu.dk⟩

FEJL

Der er rapporteret yderst få tilfælde hvor **SML/NJ** er gået ned, formentlig på grund af socket-modulet. Dette bygger vi på det grundlag af at der ikke forud for socket-modulet er stødt på fejl af denne art. Programmet kan også virke som om det "hænger" hvis en webserver er lang tid om at svare. Programmet understøtter ikke proxy forbindelser, og kører derved ikke på DIKU’s systemer.

SE OGSÅ

`uri`(7), `wget`(1), `sml`(1), `aspell`(1)

COPYRIGHT