Hjemmesideanalyse

Førsteårsprojekt

Martin Dybdal Troels Henriksen Jesper Reenberg

Datalogisk Institut, Københavns Universitet

20. juni 2007

Agenda

Problemstilling

- Hjemmesider forfattet uden omtanke for læsbarhed

Problemstilling

- Hjemmesider forfattet uden omtanke for læsbarhed
- Hjemmesider forfattet med primitive værktøjer
  - Ingen stavekontrol
  - Ingen grammatikkontrol

Problemstilling

- Hjemmesider forfattet uden omtanke for læsbarhed
- Hjemmesider forfattet med primitive værktøjer
  - Ingen stavekontrol
  - Ingen grammatikkontrol
- Store websteder kan være uoverskuelige at læsbarheds-vurdere manuelt

Målgruppe

- Webmastere

Målgruppe

- Webmastere
- Hjemmeside–skribenter

Målgruppe

- Webmastere
- Hjemmeside–skribenter
- Generelt, HTML-kyndige

Målgruppens behov

- Hele websteder skal analyseres

Målgruppens behov

- Hele websteder skal analyseres
- Skal være let at finde læsbarhedsproblemer

Målgruppens behov

- Hele websteder skal analyseres
- Skal være let at finde læsbarhedsproblemer
- Store websteder, analyseresultater skal kunne deles

Målgruppens behov

- Hele websteder skal analyseres
- Skal være let at finde læsbarhedsproblemer
- Store websteder, analyseresultater skal kunne deles
- Automatiseret programkørsel (via `cron` mfl.)

Målgruppens behov

- Hele websteder skal analyseres
- Skal være let at finde læsbarhedsproblemer
- Store websteder, analyseresultater skal kunne deles
- Automatiseret programkørsel (via `cron` mfl.)
- Skal være muligt at finde problematiske undersider på store websteder

Målgruppens behov

- Hele websteder skal analyseres
- Skal være let at finde læsbarhedsproblemer
- Store websteder, analyseresultater skal kunne deles
- Automatiseret programkørsel (via `cron` mfl.)
- Skal være muligt at finde problematiske undersider på store websteder
  - Hver side skal tildeles en enkelt talværdi der angiver dens læsbarhed

Præsentationsform

- HTML–dokumenter
  - Kan deles.

Præsentationsform

- HTML–dokumenter
  - Kan deles.
  - Intet behov for specifikt fremvisningsprogram.

Præsentationsform

- HTML–dokumenter
  - Kan deles.
  - Intet behov for specifikt fremvisningsprogram.
  - Webapplikation kan bygges ovenpå.

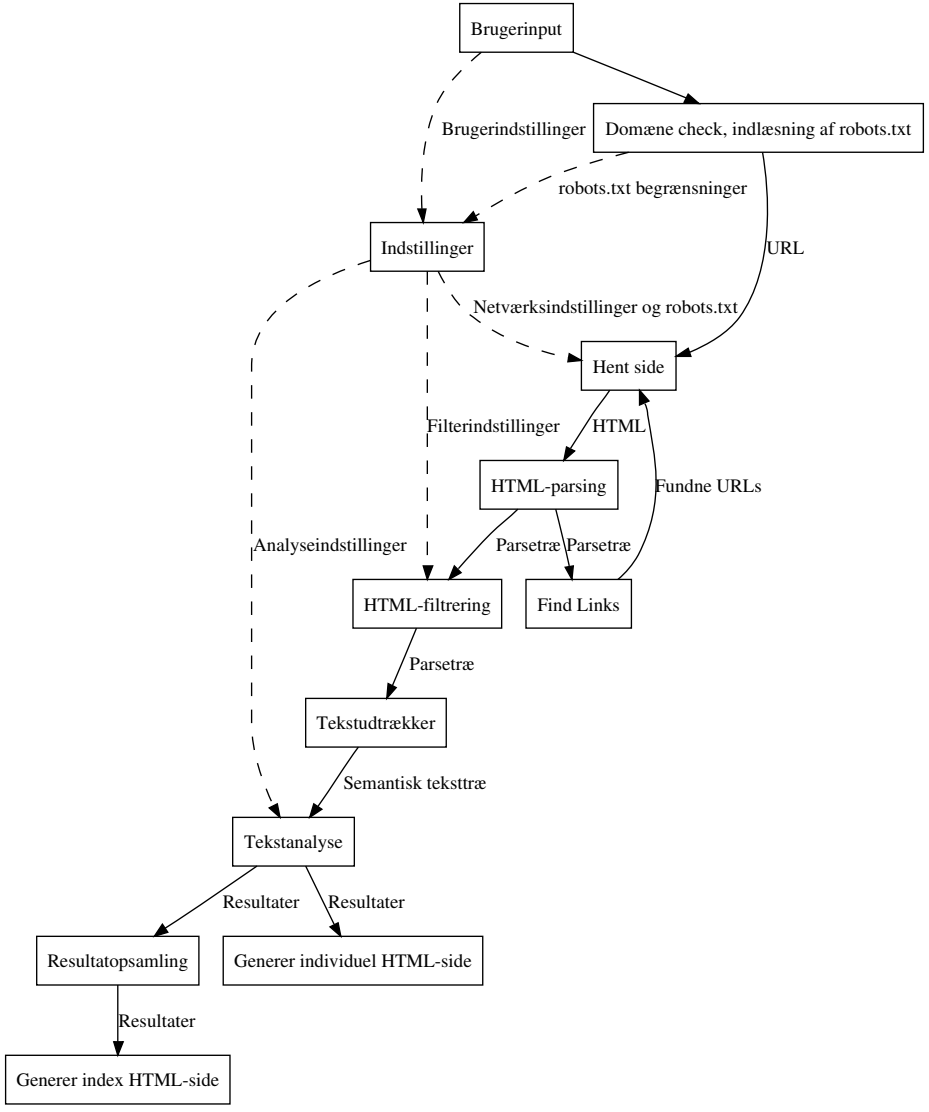
Præsentationsform

- HTML–dokumenter
  - Kan deles.
  - Intet behov for specifikt fremvisningsprogram.
  - Webapplikation kan bygges ovenpå.
- Kommandolinjeprogram
  - Let at automatisere.

Præsentationsform

- HTML–dokumenter
  - Kan deles.
  - Intet behov for specifikt fremvisningsprogram.
  - Webapplikation kan bygges ovenpå.
- Kommandolinjeprogram
  - Let at automatisere.
  - ~~Let~~ overkommeligt at skrive grafisk interface.

Dataflow


Brugerinput
Brugerindstillinger
Domæne check, indlæsning af robots.txt
robots.txt begrænsninger
Indstillinger
URL
Netværksindstillinger og robots.txt
Hent side
Filterindstillinger
HTML
HTML-parsing
Fundne URLs
Analyseindstillinger
Parsetræ
Parsetræ
HTML-filtrering
Find Links
Parsetræ
Tekstudtrækker
Semantisk teksttræ
Tekstanalyse
Resultater
Resultater
Resultatopsamling
Generer individuel HTML-side
Resultater
Generer index HTML-side

HTML–Parsing

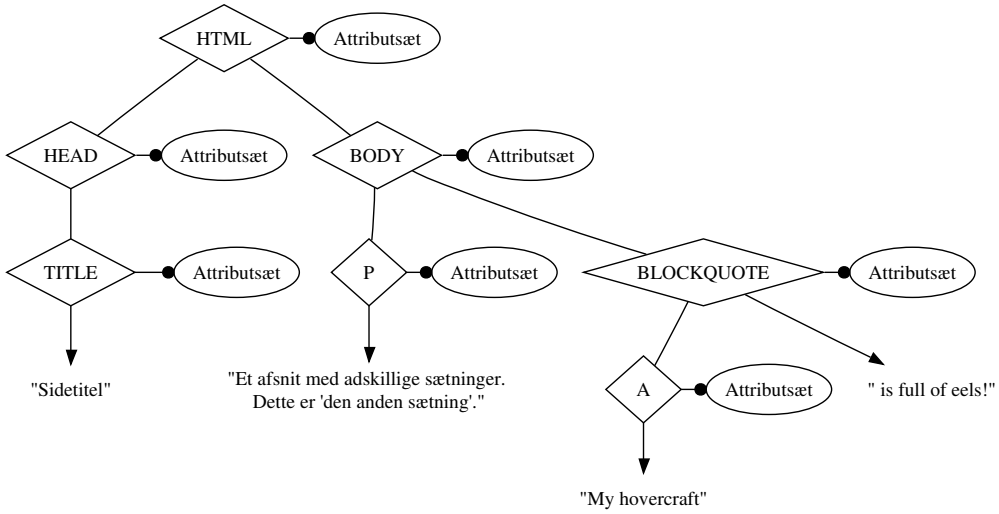
```
<HTML>
  <HEAD>
    <TITLE>Sidetitel</TITLE>
  </HEAD>
  <!-- En kommentar -->
  <BODY>
    <P>
     Et afsnit med adskillige sætninger.
     Dette er 'den anden sætning'.
     </P>
    <BLOCKQUOTE lang="en">
     <A href="http://en.wikipedia.org/wiki/Hovercraft">
     My hovercraft</A> is full of eels!
    </BLOCKQUOTE>
  </BODY>
</HTML>
```

Figur: HTML–dokument

HTML–Parsing

```
<HTML>
  <HEAD>
    <TITLE>Sidetitel</TITLE>
  </HEAD>
  <!-- En kommentar -->
  <BODY>
    <P>
     Et afsnit med adskillige sætninger.
     Dette er 'den anden sætning'.
     </P>
    <BLOCKQUOTE lang="en">
     <A href="http://en.wikipedia.org/wi
     My hovercraft</A> is full of eels!
    </BLOCKQUOTE>
  </BODY>
</HTML>
```

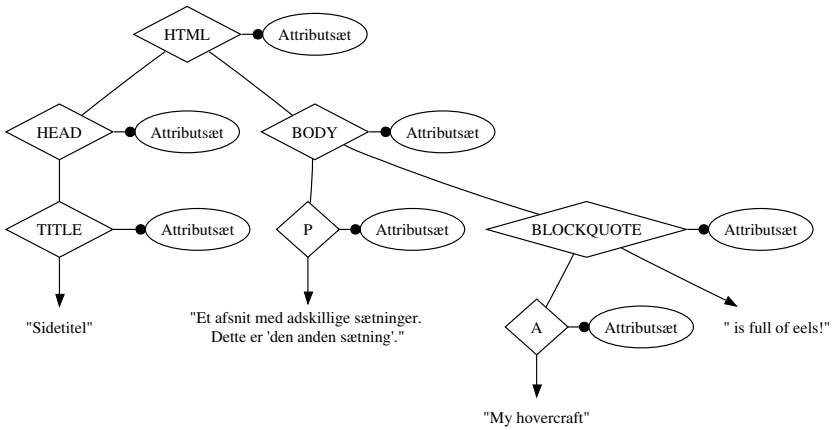
Figur: HTML–dokument



Figur: Parsetræ

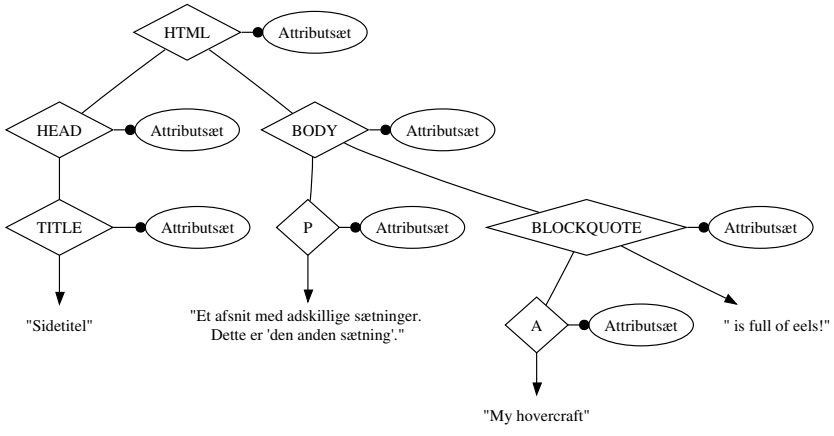
- Beholder struktur
- Beholder data

Udtræk af tekst

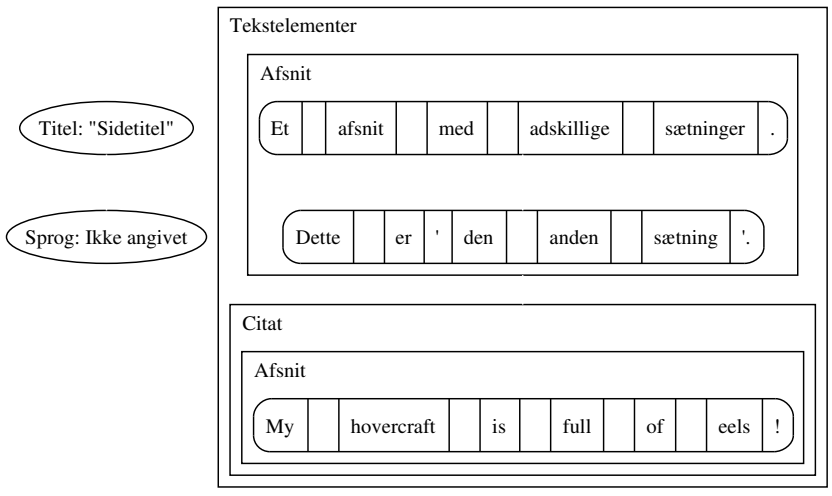


Figur: Parsetræ

Udtræk af tekst



Figur: Parsetræ



Figur: Dokument–struktur

Tekstanalyse

Analysemetoder

Beregning af Læsbarhedsindeks

$$\frac{\text{antal ord}}{\text{antal sætninger}} + 100\left(\frac{\text{antal lange ord}}{\text{antal ord}}\right)$$

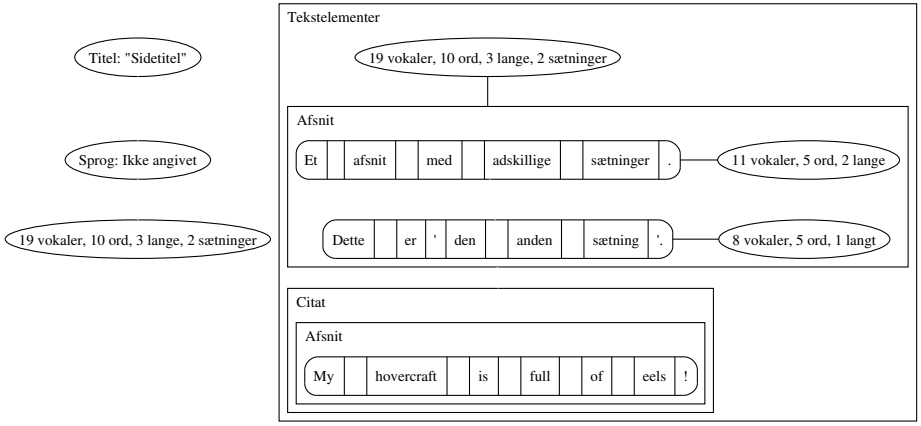
Beregning af Flesch Reading Ease

$$206.835 - 1.015\left(\frac{\text{antal ord}}{\text{antal sætninger}}\right) - 84.6\left(\frac{\text{antal stavelser}}{\text{antal ord}}\right)$$

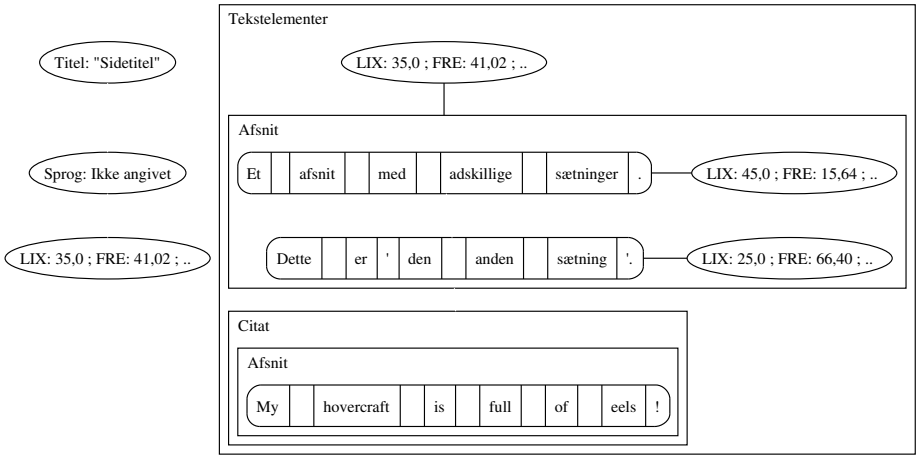
Beregning af Flesch–Kincaid Grade Level

$$0.39\left(\frac{\text{antal ord}}{\text{antal sætninger}}\right) + 11.8\left(\frac{\text{antal stavelser}}{\text{antal ord}}\right) - 15.59$$

Tekstanalyse



Figur: Optælling



Figur: Analyseresultat

Demonstration

Lix: 41.24
Flesh Reading Ease: 41.50
FK Grade Level: 12.69

Det siges, at Windows XP ikke accepterer at blive bootet fra en disk, som der er ændret på med programmer, som ikke er fra Microsoft, fx. Ghost, Novell, Partitionmagic - og Linux LILO. Efter vores viden er det en bug i MS-XP, som bevirker, at endog Microsoft setup-programmer kan lave rav i licens-kontrollen, således at man ikke får lov at starte MS-XP. Da dette selvfølgelig kan være alvorligt belastende, må vi råde til forsigtighed med de nyeste versioner af Microsoft produkter.

Lix: 31.67
Flesh Reading Ease: 53.43
FK Grade Level: 9.53

Hvis man vil bevare en MS-XP installation, må vi derfor anbefale den forsigtige metode. Vi kommenterer i øvrigt ikke på fejl i MS-XP eller hvordan man undgår dem.

Lix: 50.00
Flesh Reading Ease: 46.95
FK Grade Level: 11.68

De ældre Microsoft systemer, NT-4.x m.v., opfører sig helt korrekt og er non-destruktive overfor egne og andres data.

Lix: 44.31
Flesh Reading Ease: 18.75
FK Grade Level: 14.77

Den forsigtige metode består i, at man installerer Linux, men ikke installerer en LILO boot sector. Man benytter så den start-diskette, som man kan lave til sidst i installationen. Vi beskriver også, hvordan man selv kan lave sådan en diskette med sin egen kernekonfiguration. Det er i virkeligheden ret effektivt, idet disketten efter indlæsning af kernen sørger for, at harddisken bruges som udgangspunkt for alle andre operationer, herunder kernemoduler, hvis man benytter sådanne. Det drejer sig højst om 10-20 sekunder ekstra i opstarten.

Funktionstest

Skal afsløre...

- ...uopfyldte krav.
- ...funktionalitet der ikke virker korrekt.

Test kan ikke automatiseres, pga. uddatas format.

Funktionstest

Fundne fejl

- Det *uundværlige* krav om håndtering af robots.txt (1.1), er kun delvist implementeret. Det er ikke muligt at slå robots.txt–behandlingen fra.
- **De resterende *uundværlige* og *vigtige* krav er implementeret korrekt.**
- Det *mindre vigtige* krav (3) vedr. analysebaseret på HTML–tags er delvist implementeret.
- Det *mindre vigtige* krav (4) om konfiguration af analyse er delvist implementeret. Filtrering på basis af `class`–attributer virker ikke.

Brugertest

Krav til forsøgsperson

- Skal kende til elementær HTML.
- Skal kende til GNU/Linux, vores målplatformen.
- Skal have erfaring med kommandolinje programmer.

Brugertest

Krav til forsøgsperson

- Skal kende til elementær HTML.
- Skal kende til GNU/Linux, vores målplatformen.
- Skal have erfaring med kommandolinje programmer.

Udførelse af test

- Udført som tænk–højt forsøg.
- Vi valgte en person fra DIKU, da de fleste på DIKU har de 3 ovenfor nævnte egenskaber.
- Vi stillede ham 3 opgaver og gav ham brugermanualen.

Brugertest

Resultater

- Det er svært at se om programmet arbejder, om det er færdigt eller er gået i stå.
- Der mangler beskrivende tekst på indeks–siden, det er svært at se hvad farverne angiver og hvordan siderne er sorteret.
- Manualen var tvetydig omkring hvordan den maksimale dybde skal angives.

Konklusion

Status

- Alle vigtige krav opfyldt

Konklusion

Status

- Alle vigtige krav opfyldt
- Kun få krav ikke implementeret pga. tidsnød

Konklusion

Status

- Alle vigtige krav opfyldt
- Kun få krav ikke implementeret pga. tidsnød
- Kan udvides med flere analysemetoder og understøttelse af andre tekstformater.

Konklusion

Status

- Alle vigtige krav opfyldt
- Kun få krav ikke implementeret pga. tidsnød
- Kan udvides med flere analysemetoder og understøttelse af andre tekstformater.
- Programmet er “brugbart”

Konklusion

Status

- Alle vigtige krav opfyldt
- Kun få krav ikke implementeret pga. tidsnød
- Kan udvides med flere analysemetoder og understøttelse af andre tekstformater.
- Programmet er “brugbart”

Problemer

- Dårlig håndtering af tekstindkodning (indbefattet i krav)

Konklusion

Status

- Alle vigtige krav opfyldt
- Kun få krav ikke implementeret pga. tidsnød
- Kan udvides med flere analysemetoder og understøttelse af andre tekstformater.
- Programmet er “brugbart”

Problemer

- Dårlig håndtering af tekstindkodning (indbefattet i krav)
- Sætningsopdeler ikke helt perfekt
  - Kan let narres af forkortelser fulgt af navne
  - Ingen egentlig dybdegående grammatisk forståelse
  - Håndtering af ikke–angliske/nordiske sprog er mangelfuld

Konklusion

Status

- Alle vigtige krav opfyldt
- Kun få krav ikke implementeret pga. tidsnød
- Kan udvides med flere analysemetoder og understøttelse af andre tekstformater.
- Programmet er “brugbart”

Problemer

- Dårlig håndtering af tekstindkodning (indbefattet i krav)
- Sætningsopdeler ikke helt perfekt
  - Kan let narres af forkortelser fulgt af navne
  - Ingen egentlig dybdegående grammatisk forståelse
  - Håndtering af ikke–angliske/nordiske sprog er mangelfuld
- Respekt for `robots.txt` kan ikke slås fra
- Indeks–siden mangler forklaring af farver

Perspektiver

Mulige udvidelser

- Reel grammatisk forståelse baseret på det angivne sprog

Perspektiver

Mulige udvidelser

- Reel grammatisk forståelse baseret på det angivne sprog
- Flere analysemetoder, brugt/kombineret på mere intelligent måde

Perspektiver

Mulige udvidelser

- Reel grammatisk forståelse baseret på det angivne sprog
- Flere analysemetoder, brugt/kombineret på mere intelligent måde
- Anden brugergrænseflade — webapplikation eller grafisk applikation

Perspektiver

Mulige udvidelser

- Reel grammatisk forståelse baseret på det angivne sprog
- Flere analysemetoder, brugt/kombineret på mere intelligent måde
- Anden brugergrænseflade — webapplikation eller grafisk applikation
- Læsbarhedsstatistik for websider — gennemsnit, median, afvigelse, osv.

Perspektiver

Mulige udvidelser

- Reel grammatisk forståelse baseret på det angivne sprog
- Flere analysemetoder, brugt/kombineret på mere intelligent måde
- Anden brugergrænseflade — webapplikation eller grafisk applikation
- Læsbarhedsstatistik for websider — gennemsnit, median, afvigelse, osv.
- Flertrådet program