

Webcrawling til analyse af hjemmesider

Afgangsprojekt udført hos Morning Train Technologies ApS

Macintosh HD:Users:rubatharisan:Downloads:morning-train-logo-left-black-big.png

**Udarbejdet af:**

Rubatharisan Thirumathyam

[ruthi12@student.sdu.dk](mailto:Ruthi12@student.sdu.dk)

[rt@morningtrain.dk](mailto:rt@morningtrain.dk)

Eksamens nr.: 324539

**20. maj 2017**

**Vejledt af:**

Ole Dolriis **(SDU)** [od@mmmi.sdu.dk](mailto:od@mmmi.sdu.dk)

Bjarne Bonde **(Morning Train)** [bb@morningtrain.dk](mailto:bb@morningtrain.dk)

**Synopsis**

**Forord**

Dette projekt er udarbejdet som et afslutningsprojekt, på uddannelsen som diplomingeniør indenfor Information- og kommunikationsteknologi. Projektet er udarbejdet i perioden 01-02-2017 til 22-05-2017.

Projektet er lavet i samarbejde med Morning Train Technologies ApS (ref. Morning Train), hvor jeg har i fællesskab med Bjarne Bonde (CTO v/Morning Train) defineret projektets vision. Dette projekt er blevet vejledt af Ole Dolriis (lektor v/SDU) og Bjarne Bonde. Under projektforløbet har jeg fået råd fra alle konsulenter fra webbureaet Morning Train. Den primære interessegruppe for dette projekt er webbureauer og webudviklere.

Som led i mit uddannelsesforløb har jeg haft et praktikophold på 6 måneder hos Morning Train. Morning Train ledelsen har herefter valgt at ansatte mig som serveransvarlig, hvor mit primære ansvarsområde har været serveropsætning, servervedligeholdelse og serversikkerhed. Mit sekundær ansvarsområde har været fullstack udvikling af systemer. Da jeg startede på dette projekt, havde jeg allerede været ansat i virksomheden i 1 år og 7 måneder.

Tak til Morning Train for at stille ressourcer til rådighed så dette projekt kunne blive et realitet. Specielt tak til Bjarne Bonde og Ole Dolriis for at have vejledt mig igennem de gode og de dårlige perioder i projektet.

- Rubatharisan Thirumathyam

ruthi12, eksamens nr. 324539

**Læsevejledning**

Denne rapport er struktureret og skrevet så det giver læseren en fornemmelse af hvordan processen har været for at gennemføre dette afgangsprojekt.

Afgangsprojektet indledes i kapitel 1 med en indledning hvor jeg præsentere projektets problemområde og problemformulering. projekt er udarbejdet som et afslutningsprojekt, på uddannelsen som diplomingeniør indenfor Information- og kommunikationsteknologi. Projektet er udarbejdet i perioden 01-02-2017 til 22-05-2017.

**Indhold**

1. Indledning 6

1.1 Problemområde 6

1.2 Problemformulering 7

1.3 Problemafgrænsning 9

1.4 Læsevejledning 9

2. Planlægning 9

2.1 Udviklingsmetode 9

2.2 Agil softwareudvikling 9

2.3 Scrum 9

2.2 Projektplan 10

3. Research 12

3.1 … 12

3.2 … 12

4. Design 13

4.1 … 13

4.2 … 13

5. Implementation og test 14

5.1 … 14

5.2 … 14

6. Evaluering 15

6.1 … 15

6.2 … 15

7. Konklusion 16

7.1 … 16

7.2 … 16

8. Referencer 17

8.1 … 17

8.2 … 17

9. Appendiks 18

9.1 … 18

9.2 … 18

# 1. Indledning

Målet med dette projekt er at anvende min viden, kompetencer og færdigheder tilegnet gennem min uddannelse. Dette projekt skulle være baseret på et problematik i et virksomhed, hvorved skulle jeg anvende min viden, kompetencer og færdigheder til at fremdrage en løsning. På grund af mine rolle Morning Train, fik jeg lov til at medvirke til at danne en problemformulering, hvor jeg ville kunne anvende mine færdigheder.

## 1.1 Problemområde

Morning Train er et webbureau lokaliseret i Odense, drevet af kvalitetsbevidste webnørder. Morning Train specialisere sig primært indenfor tre kerne områder: digital design, intelligente webløsninger og online markedsføring.

Indenfor disse tre kerneområder er der mange problematiske udfordringer, som hver kunne være et projekt i sig selv. Jeg søgte et projekt som strækkede sig over alle tre kerne områder, et projekt der ville kunne være til gavn for alle tre kerne områder. Men for at kunne gøre dette, blev jeg nød til at finde ud af præcis hvordan de forskellige afdelinger fungere.

Kort beskrevet så når Morning Train udvikler en ny hjemmeside til en kunde, så er der tre afdelinger som tager hånd om opgaven. Morning Train har front-end udviklere som varetager det digitale design og layout opsætning, de har back-end udviklere der får programmering og funktionaliteten på plads. De har online markedsførings konsulenter som analysere og vurdere hjemmesiden for at efterfølgende at optimere på det de føler der vil gøre en forskel for hjemmesidens online tilværelse.

For at forstå hvad mit projekt skulle handle om, besluttede jeg at snakke med medarbejdere fra de forskellige afdelinger, for at høre dem om hvordan et afgangsprojekt ville kunne bidrage til deres afdeling. Frontend afdelingen kunne ikke rigtigt se at jeg kunne lave noget brugbart for dem, deres problemer er primært grafisk orienteret. Online markedsførings afdelingen ville have mig til at lave et system som skulle systematisk søge forskellige søgeord på Google og finde placeringer af en bestemt hjemmeside, dette projekt kunne ikke lade sig gøre, da Google ikke tillader systemer at crawle deres søgemaskine – ret banalt. Backend afdelingen havde nogle spændende projekter som alle kunne have været det perfekte projekt at bruge som afgangsprojekt. F.eks. et projekt hvor man samlede data om brugeradfærden på en hjemmeside, for at derefter bruge dette data til at forbedre hjemmesiden – hvor brugeren bliver centrum.

Men det projekt jeg syntes der vil være til gavn for alle tre afdelinger blev ”webcrawling for analyse af hjemmesider”. I en mere forstående term: en bot som systematisk besøger hver eneste underside på en hjemmeside, for at vurdere om der nogle punkter som der kan forbedres. Vurderinger af siden, kan herefter udvides, til at kunne tage hensyn til flere ting. En vurdering kunne være f.eks. en undersøgelse om alle grafiske billeder ikke overgår en bestemt filstørrelse, da større en billede er, jo længere tid skal en besøgende vente på at billedet bliver fremvist.

## 1.2 Problemformulering

En hjemmeside har en værdi for mange virksomheder, nogle virksomheder lever af deres hjemmeside. Men vi ser fejl i alle hjemmesider, nogle af disse fejl bliver aldrig opdaget, da der mange som ikke kigger specifikt efter disse specifikke fejl.

I samarbejde med Morning Train vil der blive udviklet et framework til crawling af hjemmesider. En webcrawler er en robot som systematisk besøger alle links på en hjemmeside. Man skal scripte sig op imod dette framework for at viderebehandle de data der opsamles under besøget.

Der opstilles en række testcases for at teste frameworket. Dette framework skal udvikles arkitektonisk korrekt så det er fleksibelt, skalerbart og ligetil at scripte op imod, uden at skulle koncentrere sig om skalering.

Frameworkets primære formål er at være så fleksibel at den kan understøtte en bred vifte af testcases.

Frameworket skal til slut lave en teknisk rapport som påtaler alle forbedringsmuligheder den kan finde.

Følgende testcases kan være relevant til at validere frameworkets fleksibilitet ved at belyse forskellige brugstilfælde. Testcasene er defineret på baggrund af både menneskelige og systemmæssige fejl.

Nedenstående fejl kan kategoriseres som testcases:

1. en side som ikke eksistere længere, men stadig linkes til fra en anden side på hjemmesiden. (404 fejlen).   
a. besøgende kan føle sig misledt og få et dårligere forhold til hjemmesiden.

2. en side som udløser en systemfejl. (500 – Internal server error, såkaldte http 5XX fejl)   
a. besøgende kan blive frustreret over at ikke få lov til at udføre det tiltænkte.

Weboptimerende testcases:

3. en side som har billeder der overskrider en bestemt størrelse, hvilket gør at den besøgende skal vente uretfærdigt længe på at siden bliver fremvist.   
a. den besøgende kan blive frustreret over ventetiden, og forlader hjemmesiden.

4. en side som er for svær at finde pga. den ligger for langt væk fra forsiden. (Forside -> Side A -> Side B -> Side C -> Side D)   
a. besøgende skal igennem mange sider for at komme til sit mål.

5. en side som ikke er mobil-, tablet- eller desktopvenlig (responsiv).   
a. besøgende vælger at forlade siden, pga. elementer på hjemmesiden som ikke er optimeret til dens enhed.

## 1.3 Problemafgrænsning

Da det ikke er umiddelbart muligt at estimere hvor langtid dette projekt kommer til at tage, og at jeg har en deadline til omkring 22. maj 2017, har vi som udgangspunkt besluttet at jeg kun skal lave frameworket til at kunne understøtte de første to test cases (punkt 1. og 2. fra problemformuleringen). Hvis det viser sig at projektet holder sin tidsplan, vil dele af de resterende testcases blive implementeret.

# 2. Planlægning

## 2.1 Udviklingsmetode

Udviklingsmetoden som jeg har brugt i dette projekt er Scrum, årsagen til at jeg har valgt Scrum er på baggrund af dets agile tilgang til software udvikling. For at kunne forklare hvordan jeg har implementeret Scrum, bliver jeg nød til at forklare hvordan min agile tilgang har været. Min agile tilgang har været baseret på manifesten for agil softwareudvikling.

## 2.2 Agil softwareudvikling

I den agile manifest er der nogle værdier og principper der gør udvikling af software systemer nemmere og.

<http://agilemanifesto.org/>

<https://www.youtube.com/watch?v=Z9QbYZh1YXY>

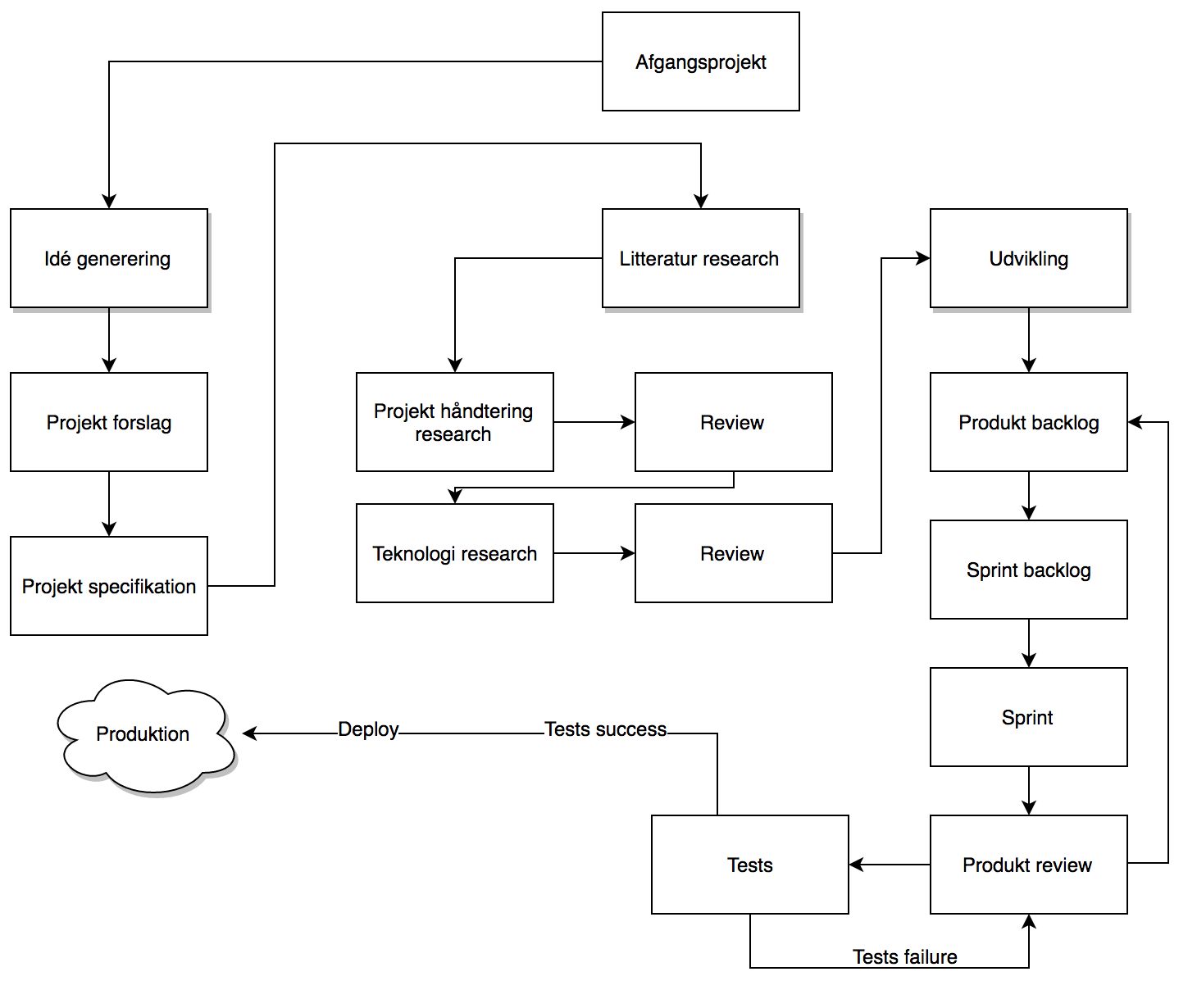
<https://en.wikipedia.org/wiki/Agile_software_development#The_Manifesto_for_Agile_Software_Development>

## 2.3 Scrum

<https://www.youtube.com/watch?v=9TycLR0TqFA>

Hvad, hvorfor.

## 2.2 Projektplan



|  |  |
| --- | --- |
| Aktivitet | Estimeret varighed |
| Idé generering | 1 uge |
| Projekt forslag | 1 uge |
| Projekt specifikation | 2 uger |
| Litteratur research | 1 uge |
| Projekt håndterings research  + Review | 2 uger |
| Teknologi research  + Review | 2 uger |
| Udvikling (opstart) | 1 uge |
| Produkt backlog (iterativt forløb) + Sprint backlog + Sprint + Produkt review + Tests + Deployment | 6 uger |

# 3. Research

## 3.1 …

## 3.2 …

# 4. Design

## 4.1 …

## 4.2 …

# 5. Implementation og test

## 5.1 …

## 5.2 …

# 6. Evaluering

## 6.1 …

## 6.2 …

# 7. Konklusion

## 7.1 …

## 7.2 …

# 8. Referencer

## 8.1 …

## 8.2 …

# 9. Appendiks

## 9.1 …

## 9.2 …