Hochschule Luzern

PAWI

Evaluation of different content extraction algorithms

Author:
Joel Rolli

Supervisor:
Patrick Huber / Patrik Lengacher

A thesis submitted in fulfilment of the requirements for the degree of some HSLU degree

in the

Research Group Name
Department or School Name

September 2014

Declaration of Authorship

I, Joel Rolli, declare that this thesis titled, 'Evaluation of different content extraction algorithms' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

HSLU

Abstract

Faculty Name
Department or School Name

some HSLU degree

Evaluation of different content extraction algorithms

by Joel Rolli

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

Contents v

List of Figures viii

List of Tables ix

Abbreviations x

Physical Constants xi

Symbols xii

1 Planning 1

1.1 Planning concept . . . 1
1.2 Milestones overview . . . 2
1.3 Delivery objects . . . 3
1.4 Milestone one - m1 . . . 4
1.4.1 Stories m1 . . . 4
1.5 Milestone two - m2 . . . 6
1.5.1 Stories m2 . . . 6
1.6 Milestone three - m3 . . . 7
1.6.1 Stories m3 . . . 7
1.7 Milestone four - m4 . . . 8
1.7.1 Stories m4 . . . 8
1.8 Milestone five - m5 . . . 8
1.8.1 Stories m5 . . . 9

A Appendix Title Here **10**

Bibliography **11**

List of Figures

List of Tables

Abbreviations

LAH **L**ist **A**bbreviations **H**ere

Physical Constants

Speed of Light $c = 2.997\ 924\ 58 \times 10^8\ \mathrm{ms}^{-\mathrm{S}}$ (exact)

Symbols

a	distance	m
P	power	W (Js^{-1})
ω	angular frequency	rads^{-1}

For/Dedicated to/To my...

Chapter 1

Planning

1.1 Planning concept

For the project planning a combination of the two well known planning frameworks scrum and RUP are used.
For a first rough planning the assignment is split into working packages and assigned to milestones. For each milestone delivery objects are defined.
This plan is then assigned to the given time table of about 12 weeks. The project effort is defined as 180 hours. This results in about 15h work load per week.

A more detailed planning is done for the incoming milestone / sprint. The predefined working packages is split into smaller packages. For the first draft, only the first milestone is split into smaller packages. The later milestones are going to be defined in more detail as soon as all needed information is available.

The effort needed for the documentation is not listed separately. All the tasks already contain additional time for updating the documentation.

The milestones dates are not finally defined. The meeting dates can vary by up to some days.

1.2 Milestones overview

Name	**Shortcut**	**Weeks**	**Estimated hours**	**Hours total**	**Closing date**
Milestone one	m1	2.5	39	39	01.10.2014
Milestone two	m2	3	45	84	22.10.2014
Milestone three	m3	2	30	114	05.11.2014
Milestone four	m4	2	30	144	19.11.2014
Milestone five	m5	2.5	38	182	08.12.2014

1.3 Delivery objects

Milestone	**Delivery objects**
Milestone one	• System specification • Sketch software architecture • Short presentation CI environment • Draft risk evaluation
Milestone two	• Elaborated software architecture • Tested code of test framework (tbd: which components) • Interface definition for justext/boilerplate components • HTML test data
Milestone three	• Working test environment with both justext and boilerplate components integrated
Milestone four	• Evaluation environment for output data of test framework • First approach for new algorithm
Milestone five	• Implementation of new algorithm • Final documentation • Final presentation

1.4 Milestone one - m1

- Closing date date: 1.10.2014
- Available time: ca. 39h

Story	**Shortcut**	**Estimated time**
Planning	s1	4h
Research HTML / Algorithms	s2	8h
System specification	s3	12h
Risk evaluation	s4	3h
Draft software architecture	s5	8h
Configuration CI environment	s6	4h
Total		39h

1.4.1 Stories m1

Title	Planning
Id	s0
Estimated time	4h
Description	As a project owner, you want to have a time schedule, when you are going to see which results. The PAWI project is split into several working packages and are split into single stories. The working packages are assignment to milestones and for each milestone, delivery objects are defined. This can be a document, a pice of test or production code or some other kind of work.

Title	Research HTML / Algorithms
Id	s1
Estimated time	8h
Description	My knowledge about HTML and content extraction algorithms is still limited. To get an idea, what I'm going to face and what I have to take in consideration for performing the first tasks, a short research on these topics is needed.

Title	System specification
Id	s2
Estimated time	12h
Description	The PAWI project is defined through a short project description. This description does not cover all necessary information to plan and perform this project. The key features, interfaces and delivered objects have to be defined more close. The system specification should cover all this requirements.

Title	Draft software architecture
Id	s3
Estimated time	8h
Description	A first rough software architecture should be made as soon as possible. This should uncover any misunderstandings between tutors and student. Further, it is much easier to plan the further steps when the software is split into several parts.

Title	Risk evaluation
Id	s4
Estimated time	8h
Description	With the gathered knowledge by defining the specification and the software architecture potential risks should be uncovered and further actions can be defined to minimize these risks.

Title	Configuration CI environment
Id	s5
Estimated time	4h
Description	To deliver high quality software a continuous integration environment is needed. Following tools should be evaluated and configured for further use. • Version control (git) • Project build automation tool (gradle) • continuous integration service (Travis CI)

1.5 Milestone two - m2

- Closing date date: 22.10.2014
- Available time: ca. 45h

Story	**Shortcut**	**Estimated time**
Implementation test framework	s6	20h
Prototype Integration of justext/boilerpipe	s7	17h
Collection test data	s8	8h
Total		45h

1.5.1 Stories m2

Title	Implementation Testframework
Id	s6
Estimated time	20h
Description	Implementation of a first part of the test framework. This story will be divided into smaller stories as soon as the software architecture and the system specification is reviewed.

Title	Prototype Integration of justext/boilerpipe
Id	s7
Estimated time	4h
Description	Implementation of a small prototype which uses the existing implementation of justext and boilerpipe. A final interface for both components needs to be defined for further use. This story will be divided into smaller stories as soon as the software architecture and the system specification is reviewed.

Title	Collection of test data
Id	s8
Estimated time	8h
Description	To evaluate the functionality of the text extraction algorithms a certain amount of test data is needed. This test data contains HTML files of several web pages. The HTML code is categorized into content and boilerplate.

1.6 Milestone three - m3

- Closing date date: 5.11.2014
- Available time: ca. 30

Story	**Shortcut**	**Estimated time**
Implementation test framework	s9	20h
Final integration of justext / boilerplate	s10	10h
Total		30h

1.6.1 Stories m3

Title	Implementation test framework
Id	s9
Estimated time	20h
Description	Final implementation of the test framework. This story will be divided into smaller stories as soon as the software architecture and the system specification is reviewed.

Title	Prototype Integration of justext/boilerpipe
Id	s10
Estimated time	4h
Description	Complete integration of the justext and boilerplate algorithms into the test framework. This story will be divided into smaller stories as soon as the software architecture and the system specification is reviewed.

1.7 Milestone four - m4

- Closing date date: 19.11.2014
- Available time: ca. 30h

Story	**Shortcut**	**Estimated time**
Evaluation environment for results	s11	20h
Research on new algorithm	s12	10h
Total		30h

1.7.1 Stories m4

Title	Evaluation environment of results
Id	s11
Estimated time	20h
Description	The test framework will produce a lot of output data. To review this data an evaluation environment is needed which should process this data and present the results in a descriptive way. This story will be divided into smaller stories as soon as the software architecture and the system specification is reviewed.

Title	Research on new algorithm
Id	s12
Estimated time	20h
Description	A first research on the new algorithm should be performed. After this research it should be possible to decide if this solution is possible and if an implementation with the remaining time resources is realistic. This story will be divided into smaller stories as soon as the software architecture and the system specification is reviewed.

1.8 Milestone five - m5

- Closing date date: 8.12.2014

- Available time: ca. 38h

Story	**Shortcut**	**Estimated time**
Implementation of new algorithm	s13	19h
Complete documentation	s14	15h
Prepare final presentation	s15	4h
Total		38h

1.8.1 Stories m5

Title	Implementation of new algorithm
Id	s13
Estimated time	19h
Description	Implementation of the new algorithm and analyzing the test results with the existing evaluation environment.

Title	Complete documentation
Id	s14
Estimated time	15h
Description	Complete and review all chapters of the documentation.

Title	Prepare final presentation
Id	s15
Estimated time	4h
Description	Prepare the final presentation and the final printed / digital version of the thesis.

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

[1] A. S. Arnold, J. S. Wilson, and M. G. Boshier. A simple extended-cavity diode laser. *Review of Scientific Instruments*, 69(3):1236–1239, March 1998. URL http://link.aip.org/link/?RSI/69/1236/1.

[2] Carl E. Wieman and Leo Hollberg. Using diode lasers for atomic physics. *Review of Scientific Instruments*, 62(1):1–20, January 1991. URL http://link.aip.org/link/?RSI/62/1/1.

[3] C. J. Hawthorn, K. P. Weber, and R. E. Scholten. Littrow configuration tunable external cavity diode laser with fixed direction output beam. *Review of Scientific Instruments*, 72(12):4477–4479, December 2001. URL http://link.aip.org/link/?RSI/72/4477/1.