



# IKT407 Web-mining and data analysis

## Introduction

Morten Goodwin Olsen

# Outline of today's lecture

08:00-09:00

- Web-mining and data analysis introduction
  - Examples
  - Course outline
- Presentation of projects
- Individual Introduction Round (who is who)

09:00-10:00

- Intellectual Property Rights on the World Wide Web by Bruce Perens

10:00-11:00

- Web Accessibility by Mikael Snaprud



# What will you learn?

- Pattern recognition

- How can I use basic statistics to recognise a pattern (such as patterns in the stock market – if a given stock will increase or decrease in value).

- Nouredine Bouhmala

- Example:

Two stocks Y,Z

Monitor Y,Z for some time

Stock Y has a very similar behaviour between day 6-20 as stock Z has between day 13-27.

Conclusion: Stock Z will act in day 28 as stock Y did in day 21 (with a given probability).



# What will you learn?

- Crawling / Resource allocation

- How is it possible to search the almost infinite world wide web in less than a second with a search engine such as Google.

- Morten Goodwin Olsen

- Example:

- When a web page is visited (crawled) each hyperlink is extracted.

- The next web page to visit is one of the hyperlinks just extracted.

- This simple random walk automatically crawls all web pages based on the Google pagerank.



# What will you learn?

- Text Mining / Classification / Linguistics
  - How can I use statistics to classify text on a web site into e.g. relevant / not relevant information for me or if the text is in English / Chinese / Norwegian / Spanish / German ...
  - Annika Nietzio
  - Example:
    - One website  $A$  is either in English ( $e$ ) or Norwegian ( $n$ )
    - The 1000 most common words for  $e/n$  has been stored in the dictionaries  $d(e)$  and  $d(n)$
    - By counting the number of words in  $A$  also in  $d(e)$  and in  $d(n)$ , with a very good certainty that the dictionary with most common words with  $A$  is the correct language.



# What will you learn?

- Python programming
  - How can I make applications in almost no time at all using the Python programming language compared to for example Java.  
(<http://www.python.org>)
  - Nils Ulltveit-Moe
  - Example:

```
>>> import urllib2
>>> f = urllib2.urlopen('http://www.eiao.net/webmining/news')
>>> print [line for line in lines if 'Morten Goodwin Olsen' in line
and 'e-mail' in line]
....<a href="mailto:morten.g.olsen@hia.no">Morten Goodwin
Olsen</a>...
```



# What will you learn?

- Clustering / Partitioning

- How can I divide my data into an efficient cluster.

- Nouredine Boumala

- Example:

- All records in a large distributed database are accessed in pairs of two.

- Every two records who are often accessed together will beneficially be located in the same database.

- This could increase the reading speed of the database extremely much.



# What will you learn?

- Intellectual Property Rights
  - What property rights must I be aware of when it comes to the world wide web.
  - Bruce Perens.



# What will you learn?

- Web Accessibility

- What is web accessibility and how can large scale web accessibility assessment be carried out.

- Mikael Snaprud

- Example:

- Web Content Accessibility Guidelines (WCAG) of how to design web pages.



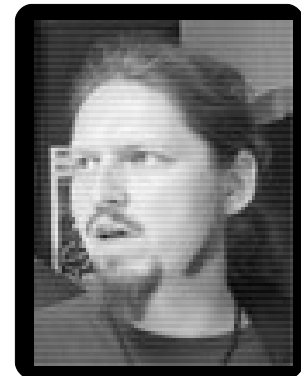
# What will you learn?

- Zope / Plone / Zope Topic Maps
  - How can I develop an advanced web application with very little effort compared to for example using Tomcat.
  - Geir Bækholt / Tor Oskar Wilhelmsen
  - Example:  
plone.org, [www.forskning.no](http://www.forskning.no), Creative Commons ([creativecommons.org](http://creativecommons.org)), Free Software Foundation ([www.fsf.org](http://www.fsf.org)), eBay ([developer.ebay.org](http://developer.ebay.org)), Lufthansa ([www.lufthansa.com](http://www.lufthansa.com)), Nasa ([www.jpl.nasa.gov](http://www.jpl.nasa.gov)), ....

# Who is part of the web-mining course?



- Nouredine Bouhmala
  - Vestfold University College
  - Classification and Clustering
- Geir Bækholt
  - Plone Solutions
  - Zope / Plone Development
- Annika Nietzio
  - Forschungsinstitut Technologie-Behindertenhilfe
  - Text mining and document classification with focus on computational linguistics



# Who is part of the web-mining course?



- Morten Goodwin Olsen

- Agder University Collage / European Internet Accessibility Observatory
- Crawling Techniques and Resource Allocation



- Bruce Perens

- Source Labs / Father of Open Source
- Intellectual Property Rights on the World Wide Web



# Who is part of the web-mining course?



- Mikael Snaprud

- Agder University College / European Internet Accessibility Observatory
- Web Accessibility



- Nils Ulltveit-Moe

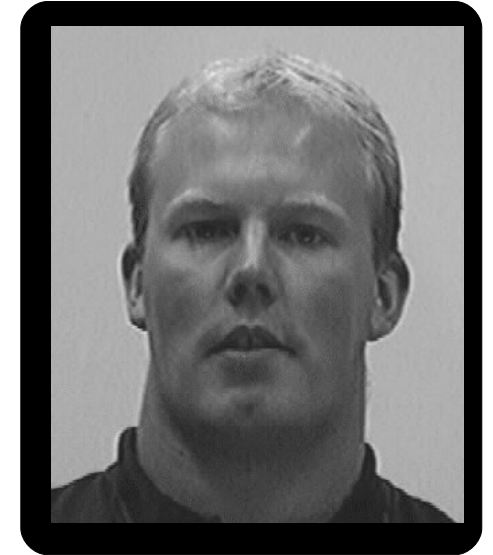
- Agder University College / European Internet Accessibility Observatory
- Python Programming



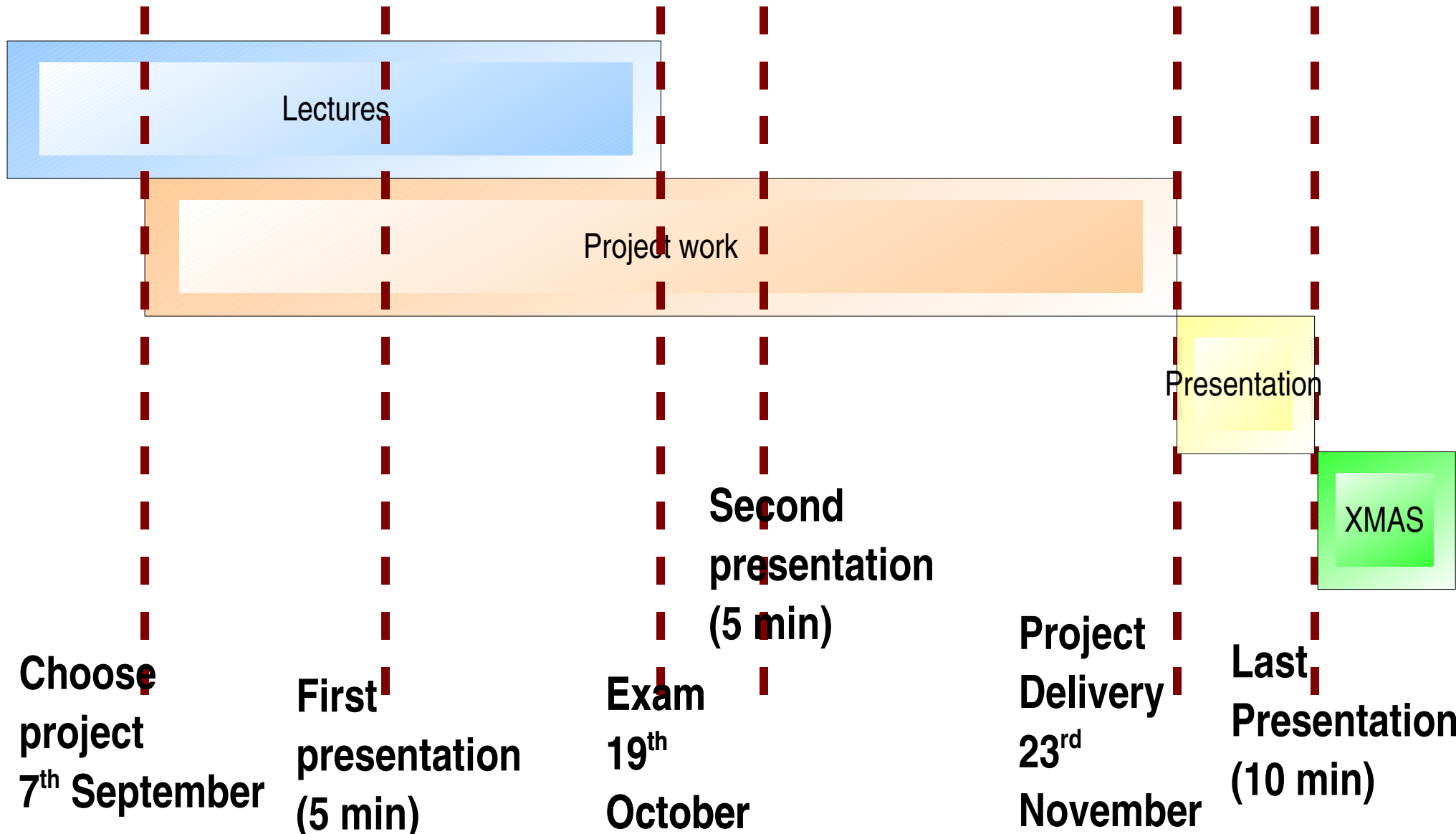
# Who is part of the web-mining course?



- Tor Oskar Wilhelmsen
  - Bouvet
  - Zope Topic Maps Development
- Leiming Chen
  - Student assistant ([leimic05@student.hia.no](mailto:leimic05@student.hia.no))



# Web-mining and data analysis schedule



# Lectures

- 2006-08-24
  - Morten Goodwin Olsen Welcome to Web-mining.
  - Bruce Perens Intellectual Property Rights on the World Wide Web
  - Mikael Snaprud Introduction to web accessibility and the European Internet Accessibility Observatory (EIAO)
- 2006-08-31
  - Nouredine Bouhmala Classification with focus on decision trees, neural network and bayesian classifiers.

# Lectures

- 2006-09-07
  - Annika Nietzo Text-mining and document classification with focus on computational linguistics.
- 2006-09-14
  - Andreas Prinz Introduction metamodelling and SMILE.
- 2006-09-21
  - Nils Ulltveit-Moe Introduction to Python programming
- 2006-09-22
  - Morten Goodwin Olsen Crawling techniques and resource allocation

# Lectures

- 2006-90-28
  - Nouredine Bouhmala Clustering
- 2006-10-05
  - Geir Bækholt Introduction Zope/Plone development
- 2006-10-06
  - Tor Oskar Wilhelmsen Zope development with focus on Zope Topic Maps (ZTM)
- 2006-10-19
  - Written examination for the students

# Projects

- Projects must be selected before 7<sup>th</sup> of September.
  - Upload to ClassFronter.
  - Names of group members.
  - At least three projects in prioritised order.
  - One project will be handed to you.

# Possible projects

- Browser based Web Accessibility Measurement Component
- The goal of this project, is to demonstrate the concept of a web browser based measurement component using Mozilla[Moz]. Mozilla is interesting as a container for WAM measurement systems due to Mozillas Accessibility architecture[MOZ-A] which is used by 3rd party software like screen readers, magnifiers, and voice dictation software, which need information about document content, UI controls and events like changes of focus.

# Possible projects

- Web Structure Mining
- The goal of this project is to construct a crawler that identifies the interpage structure of a web site. The interpage structure can be modeled e.g. as a graph. From such a graph, different quantitative measurements are to be made: e.g., link density, number of cycles in the graph, average length of paths in the graph (after cycles have been removed), and so on. Finally, it should be determined whether the navigability of a web site can be ranked meaningfully based on such measurements.

# Possible projects

- Web server log usage mining
- The goal of this project is to construct a parser that segments a web server log from a web site into sessions, and identifies the sequences of web pages have been accessed within each session. Based on such sequences, key use scenarios of the web site are to be identified. A statistical analysis of the key use scenarios identified will then be performed to identify length, variance, entry-point and exit point in term of URL's and IP's.

# Possible projects

- Web Content Mining
- The goal of this project is to create a crawler/classifier that downloads the images in a web page and tries to classify the content of each image into different categories, e.g., mathematical formula, logo, buttons, and so on. The focus should be on automatic detection of image usage that reduces the accessibility of a web page.

# Possible projects

- User Behaviour Logging for Datawarehouse Tuning
- The goal is to design a set of tools to log user behaviour for tuning the datawarehouse. The approach can be based on storing queries with timestamps and store clickstreams of the site navigation. Analysis of both sources may yield valuable information for improving the datawarehouse performance and the web interface

# Possible projects

- Authoring tool identification
- This project was a part of an ongoing research project at Agder college university named ROBACC which goal is to develop an automated internet spider that assess the accessibility of web pages. Our involvement was to develop a prototype of a classifier-module that could identify the authoring tool used to create any given webpage based on the structure of the html-code. (Should be updated or rewritten)

# Possible projects

- Accessibility scorecard in GIS system
- The goal of the project is to create a Geographical Information System based on an Open Source GIS module that can present accessibility measures as colour coded scorecards according to the scorecards defined in UWEM for the EU NUTS regions. Data to be presented may consist of HTML deviations or other data that will be extracted from the EIAO RDF repository, or possibly the EIAO datawarehouse.

# Possible projects

- Temporal web structure mining
- The goal of the project is to do an analysis of the change, growth, accessibility and interlinking of websites over time, by using The WayBack machine in combination with a web crawler. The case study should include AUC's home page.

# Possible projects

- Object Migration Automaton (OMA) for Topic Detection and Tracking In this project the students are to investigate whether a variant of the Object Migration Automaton (OMA) [Oommen and Ma, 1988] can be used for TDT. Each automaton object will be associated with an article, and a probabilistic article similarity function will be used to compare articles. The OMA seems particularly promising for TDT because it (1) learns incrementally/on-line, (2) handles noise, and (3) has low computational complexity.

# Possible projects

- Transformation between OCL and Schematron
- The goal of this project is to have a link between two constraint description languages. OCL is the OMG constraint language, which is used in the context of UML. There are several freely available OCL checking engines. Schematron is a XML constraint description language which is used in the context of web documents. The EIAO project also has a checking engine for it. Both languages are based on predicate calculus and therefore similar in expressivity. However, their appearance is not the same. The project is about finding a link between the two and implementing a transformation out of it, preferably in both directions. The link between the two languages should be expressed as a common abstract syntax or meta-model.

# Possible projects

- e-Content accessibility for dyslexic people
- Develop a WAM (Web Accessibility Metric) that can be used within EIAO (European Internet Accessibility Observatory) to produce special scores about accessibility for people with reading difficulties. The new metric can be deduced from the literature that describes the special requirements for content and layout / presentation

# Possible projects

- A multilevel Approach to K-kustering
- Data clustering is one of the common techniques in dataming. In this project, a multilevel schema is used for K-clustering problems. Multilevel techniques refer to the process of diving large and difficlut problem into smaller ones, which are hopefully much easier to handle, and then work backward towards the solution of the original problem, using a solution from a previous level as a starting solution at the next level. In this project, we introduce a combination of the multilevel paradigm with a popular algorithm used to solve the clustering problem. Large random data sets will be generated in order to judge the quality of the clustering.

# Possible projects

- A multilevel Local Search Method to K-clustering
- Supervisor: Nouredine Bouhmala
- Data clustering is one of the common techniques in dataming. In this project, a new appraoch combining a new local search method and the multilevel paradigm is introduced for solving the k-clustering mproblem. Multilevel techniques refer to the process of diving large and difficlut problem into smaller ones, which are hopefully much easier to handle, and then work backward towards the solution of the original problem, using a solution from a previous level as a starting solution at the next level. The proposed appraoch starts by coarsening the original problem into a sequence of smaller problems using coarsening scheme. Thereafter a solution to the K-clustering problem is determined at the smallest problem and is projected back to the original problems by going through a refinement pahse using local search at the each intermediate level. Large random data sets will be generated in order to judge the quality of the clustering.

# Possible projects

- Classification of web-based discussions using Naive Bayes. (three projects in total)
- Given a set of web-based discussions on various topics written in various languages, the classification problem consists of determining for each discussion (and its sub-posts) on what topic these discussions report on, and in what language they are written in. In this project the students are to investigate whether the Naive Bayes algorithm is applicable to classifying web-based discussions. The students will be given a training-set of articles and a large corpus of articles that they are to investigate on. The project will be performed in cooperation with Integrasco A/S.

# Possible projects

- Resource allocation algorithm
- Resources can be allocated everywhere, but to get the most optimal positions for all of them and waiting to be used in best efficiency, A proposed solution is to use OMA algorithm. The students will attempt to solve the problem, and the resources may have different value depending on its position. The standard for evaluating whether the resource object is the most optimal situation can be different, such like time, distance and so on. Using the Object Migration Automaton (OMA) towards partitioning the resource objects to receive the most viable solution seems like a viable approach.

# Possible projects

- Bin packing problem
- Evaluating a data analysis solution of distribution of values for the bin packing problem where the values of the Objects cannot be know before they are instantiated. The bin packing problem is defined in [BINPACK] as where "Objects of different volumes must be packed into a finite number of bins of capacity in a way that minimizes the number of bins used." This analysis can be seen as a formal examining any resource allocation problem such as a distributed crawler. In this project a solution of competitive game of learning automata should be the main focus of the distribution of the objects.

# Possible projects

- The El Farol Bar problem
- Evaluation possible solutions to the El Farel Bar problem. The El Farel Bar is defined as following from [ELFAROL]: There is a particular, finite population of people. On Thursday night, all of these people want to go to the El Farol Bar. However, the El Farol is quite small, and it's no fun to go there if it's too crowded. So much so, in fact, that the following rules are in place: - If less than 60% of the population go to the bar, they'll all have a better time than if they stayed at home. - If more than 60% of the population go to the bar, they'll all have a worse time than if they stayed at home. The interesting part of this problem, is that any deterministic solution will fail since if all chooses the same solution they will fail. Evaluation the El Farel Bar problem with a learning mechanism might give good results. This could also be mapped to real life problems such as resource allocation whenever each node does not know the situations of the other nodes. In this project a solution of competitive game of learning automata should be the main focus of the distribution of the objects.

# Possible projects

- Your suggestion.

# Individual Introduction Round

- Your name
- What do you want out of the web-mining course?
- What previous experience do you have with these topics?