

IKT 407 – Motivation and planned experiments

Group 5:

Tarjei Romtveit and Thomas H. Gøytil

Predefined problem statement

We want to design and implement a classifier, that is able to classify a discussion board type to easier identify new and unknown discussion boards by a crawler. This will also make the whole process of transforming the forums into more friendly formats than html, more automated.

Hypothesis

Given a downloaded forum html/xhtml which forum type e.g. phpbb, vbulletin, jive etc. , do this page belong to



Why is this classification so important?

- Make a manual classification task automated
- Increase the efficiency gathering unknown discussion boards to crawl

How is it manually classified today?

- Copyright/Powered by at the bottom of the page
- Metainfo
- ID and Classes
- URL structure in Links

How we will do it

- Pattern Recognition
 - Observations
 - XHTML/XQUERY -> XML -> Java object -> Remove unwanted tokens and stop words
 - Pattern Recogniser
 - Learn Naive Bayes
 - Classifier
 - Naive Bayes

Testing

- To ensure code quality we will use unit testing, where we test every method firmly
- Testing routines that will try to classify an unknown document. And make a probabilistic assumption of the accuracy of the classifier

Difficult task to overcome

- Finding out what to extract
 - Classes
 - Hrefs
 - Link to forum page
- Stop words
 - Http://
 - Domain name
 - General file extensions like ".html, .css, .js"
- Training samples
 - Getting enough data to train the classifier

Questions?