

# Web mining project

## Document Classification

Group 6

First Presentation 22.09.2005

# summary

Nowadays, internet has become more and more important and popular in the world. Accessibility of the content of Internet is mainly decided in the form of the content (html-code, colors, pictures, animations, etc.). In fact, the form of the content of Internet can determine to what extent people with disabilities can utilize the information. Assessing understanding extent of the Internet content is therefore important, and is the main purpose of our project given in this course.

Our group's project is 'document classification'. The goal of the project was to make a program to decide whether the texts are good or bad for dyslexic readers based on the text examples. The program was developed using python and mysql. The algorithm used was naive bayes.



# Work plan

- Learn to use Python and understand the naïve Bayes algorithm.
- Collect the text examples and determine metrics.
- Make a program to get the relevant probability.
- Make a classifier and test it.
- Writing the main project report and prepare presentation.



# literature survey and findings

Before we do this project, we look at some related books and websites and get some findings in the following list:

[1] Dyslexic.com - <http://dyslexic.com>

[2] Web Content Accessibility Guidelines - <http://www.w3.org/TR/WAI-WEBCONTENT/>

[3] Python Programming language – [www.python.org](http://www.python.org)

[4] The former report - <http://eiao.net/webmining/previousprojects>

[5] <DATA MINING> Ian H.Witten & Eibe Frank

**Finding one metric :**

**length of words in the web page samples.**



# problem statement

The purpose of the classifier is to decide whether a document should be assigned to a particular category or not. In our case, there are two categories:

- Well designed for dyslexic readers
- Not well designed for dyslexic readers

It is a kind of text classification, It was therefore liable to assume that methods seen applied for text-classification also could be used for our specific purpose. So we would need to base our classification on textual content. At first we should get some proper metrics, and then we need to train the necessary probabilities according to the Naïve-Bayes Algorithm which is One of the more successful and known algorithms for learning to classify text and is rather easy to implement as a good basis for the prototype of our classifier.

# e.g

Good :

4,5,6,7  $\longrightarrow$   $P(4/g)=0.5$   
 $P(5/g)=0.25$   
 $P(6/g)=0.25$

Bad:

6,7,7,6  $\longrightarrow$   $P(6/g)=0.5$   
 $P(7/g)=0.5$

Prior probability:

$P(4)=0.25$     $P(5)=0.125$     $P(6)=0.375$     $P(7)=0.5$



# Requirements in our project

- Need enough materials which can be easily distinguished from good to bad for dyslexic readers
- Master the programming skills in python
- Understand the Naïve-Bayes Algorithm and know how to use it in our program
- Require to find as many metrics as possible and choose the proper ones.
- Train the text samples and get the relevant probabilities.
- Test our results and verify our program
- Need to write a report for our project.



# Additional Content

## Roles in the project

**Kun Yang** - leader of the group. Cooperate with the other partner in distributing the work, collecting the text examples and metrics , performing training , testing result and handing in the final report

**Fei Yao** - member of the group. Analyzing the project, collecting the text examples and metrics, programming classifier and writing the webpage report.

## contact information

	<b>e-mail</b>
KunYang	<a href="mailto:kuny05@student.hia.no">kuny05@student.hia.no</a>
FeiYao	<a href="mailto:feiy05@student.hia.com">feiy05@student.hia.com</a>

<b>tel</b>
97431020

