

**Document Classification**  
**Based on**  
**Flesch**  
**Reading Ease Test**

Group 5, ICT 407  
HiA, 2006

Supervisor : Annika Nietzio  
Morten Goodwin Olsen



# Problem Description



- **Assignment**

Develop a WAM (Web Accessibility Metric) that can produce special scores about accessibility for people with dyslexia

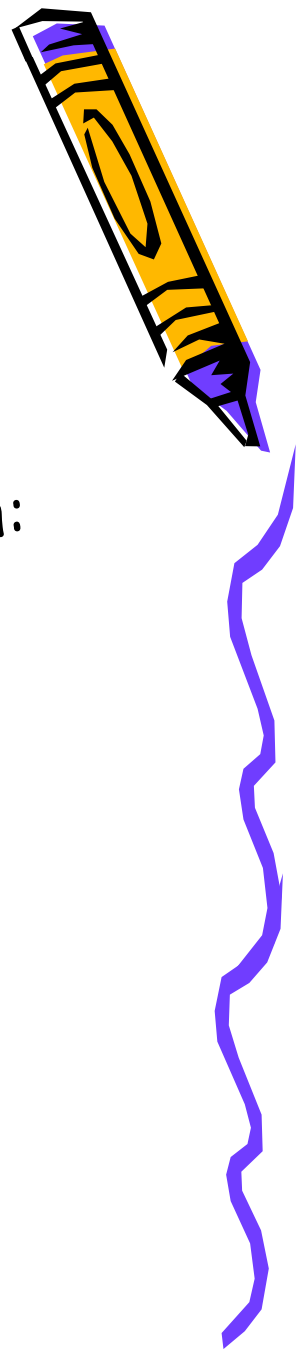


# Background

- **What is dyslexia**

People with dyslexia often have problems with:

- Visual processing (inc. scotopic sensitivity)
- Phonological decoding, analysis and processing
- Reading and comprehension
- Auditory processing
- Memory recall
- Structure and sequencing
- Planning and organization



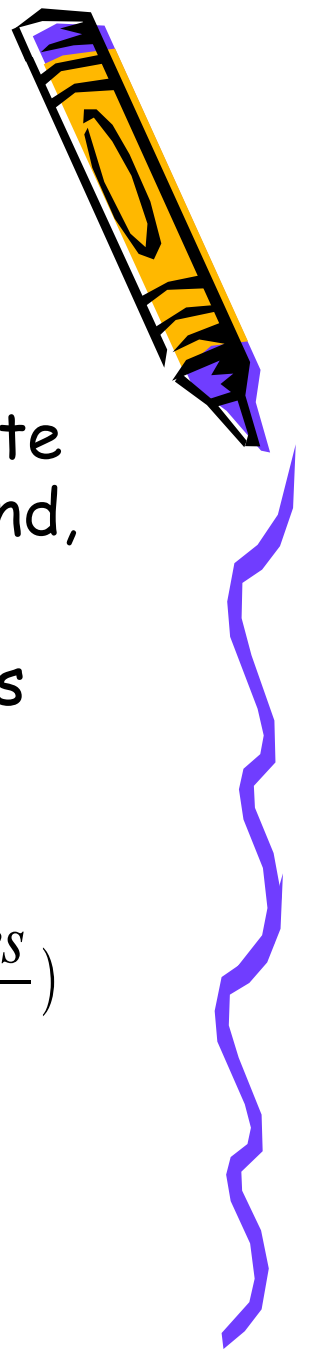
# Solution

Make a text classifier that classifies inputs into the following defined categories:

- Accessible for people with dyslexia
- Accessibility barriers exist for people with dyslexia



# Flesch Reading Ease Test



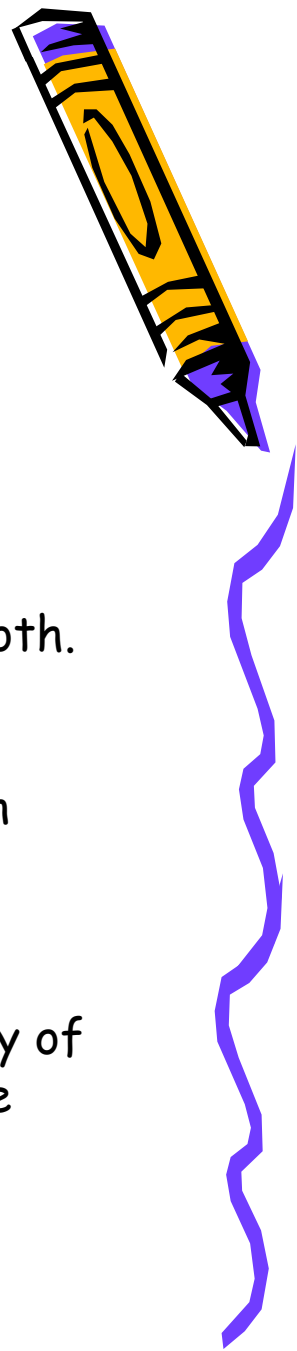
- Flesch Reading Ease test is designed to indicate how difficult a reading passage is to understand, and has become a U.S governmental standard
- In the Flesch Reading Ease test, higher scores indicate material that is easier to read; lower numbers mark harder-to-read passages:

$$206.835 - 1.015 \left( \frac{\textit{TotalWords}}{\textit{TotalSentences}} \right) - 84.6 \left( \frac{\textit{TotalSyllables}}{\textit{TotalWords}} \right)$$



# Naïve Bayes Classification

$$P(cls_i|obs) = \frac{P(obs|cls_i) * P(cls_i)}{P(obs)}$$



- Given a observation as  $obs$  of particular sample
- Suppose that either hypothesis  $cls_1$  or  $cls_2$  may occur, but not both.
- $P(cls_i)$  is the prior probability associated with hypothesis  $cls_i$
- $P(obs)$  is the probability of the occurrence of the observation
- $P(obs|cls_i)$  is the conditional probability
- $P(cls_i|obs)$  is the posterior probability, indicating the probability of occurrence of particular class by giving observation in advance



# Naïve Bayes classifier

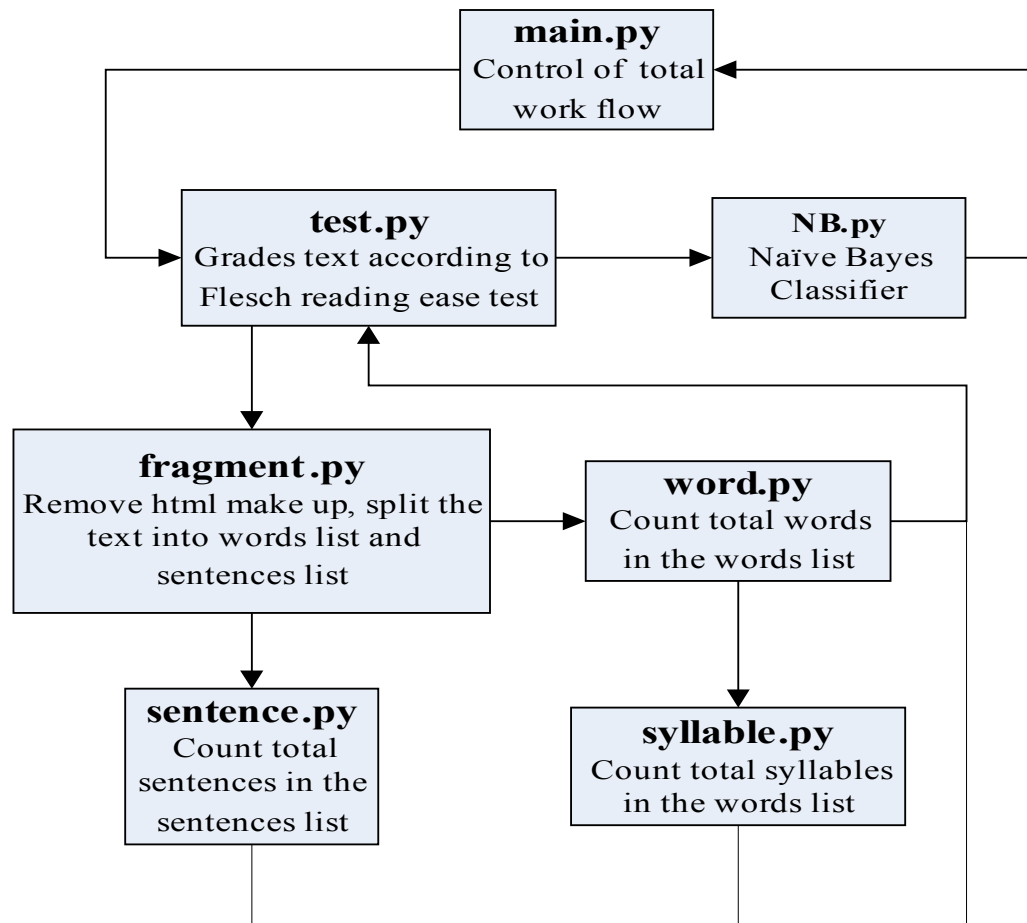
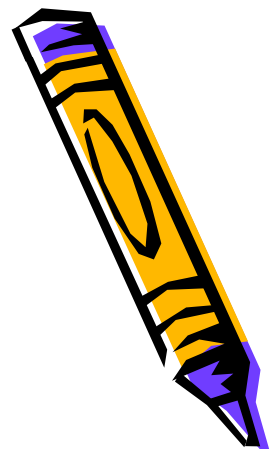
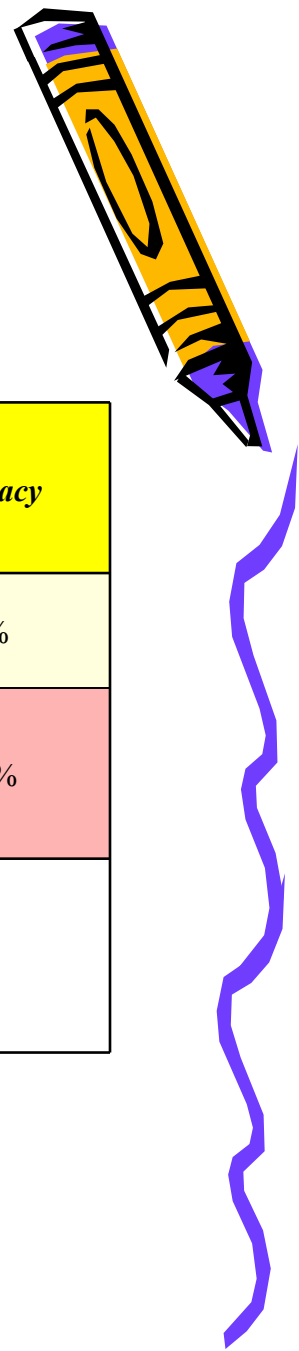


Figure 1 Architecture of modules



# Verification and testing



- Test with training data

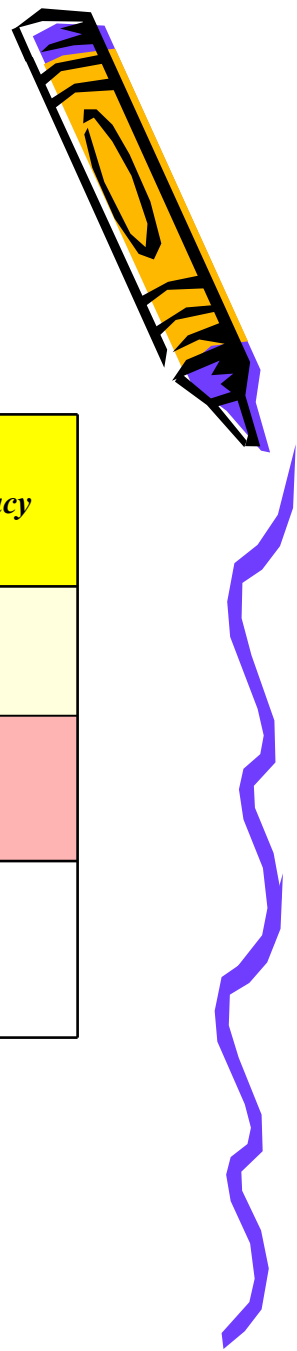
<i>Test pages Result</i>	<i>Accessible for dyslexic people</i>	<i>Not accessible for dyslexic people</i>	<i>Accuracy</i>
Accessible for dyslexic people	2	0	40%
Not accessible for dyslexic people	0	5	100%
Fail to classify according to the training dictionary	3	0	



# Verification and testing

- Test with testing data

<i>Test pages Result</i>	<i>Accessible for dyslexic people</i>	<i>Not accessible for dyslexic people</i>	<i>Accuracy</i>
Accessible for dyslexic people	0	0	0%
Not accessible for dyslexic people	1	1	50%
Fail to classify according to the training dictionary	4	4	



# Evaluation of results

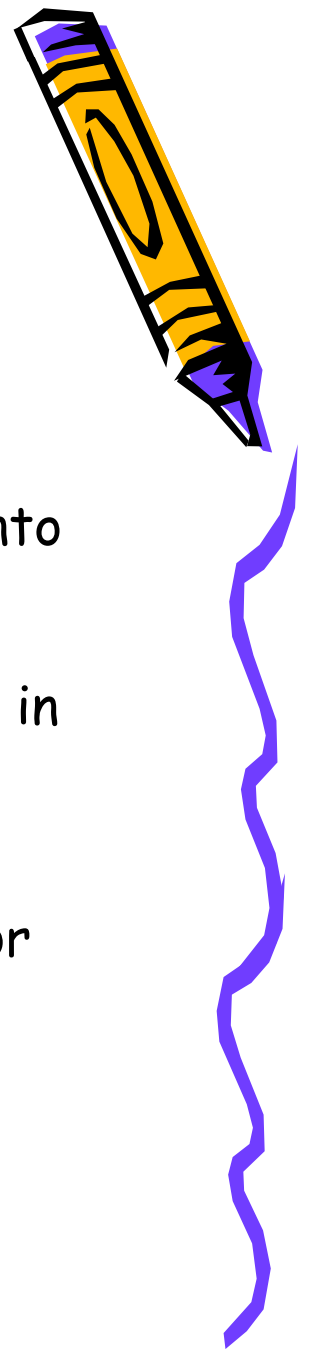


Probable reasons for this pretty low accuracy of the classifier:

- Bugs in fragment.py which removing html make up and splitting text into words list and sentences list
- Bugs in the syllable.py which estimating total syllables in the words list
- Bias in training data
- Insufficient training samples



# Further Development



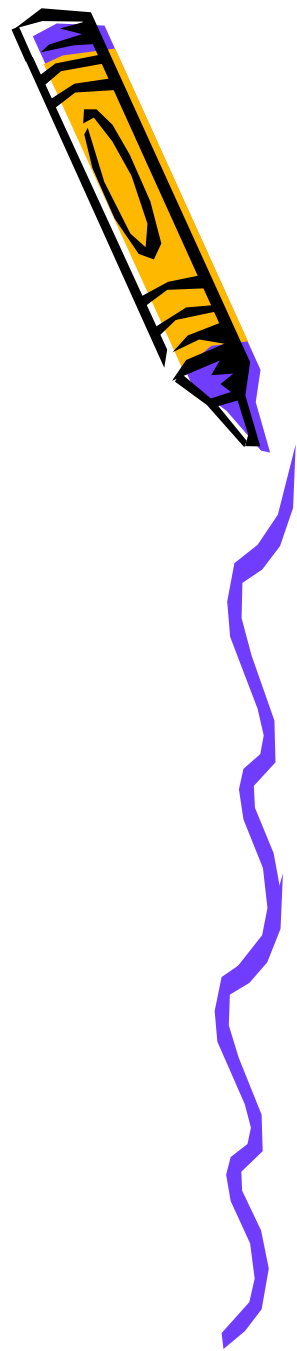
To obtain a satisfied accuracy, necessary improvements present as follows:

- Use some third party extension modules to make `fragment.py` behave better when converting html file into its plain form.
- Adopt more accurate algorithm when counting syllables in the chosen web page.
- Make an investigation with dyslexic participants, which might help distinguish web pages that are accessible for them or not.



# Further Development

- Further extensions on other languages
- Graphical User Interface (GUI) development



# Literature survey and findings



- Document Classification, Kun Yang & Fei Yao, November, 2005, HiA
- A Dyslexic Perspective on e-Content Accessibility, Peter Rainger, 20/01/03
- Naive Bayes classifier  
[http://en.wikipedia.org/wiki/Naive\\_bayes\\_classifier](http://en.wikipedia.org/wiki/Naive_bayes_classifier)
- Flesch Reading Ease test  
<http://en.wikipedia.org/wiki/Flesch-Kincaid>
- A metric measuring syllables in a word  
<http://viewvc.red-bean.com/gnoetics/src.syllables.py?view=makeup>

