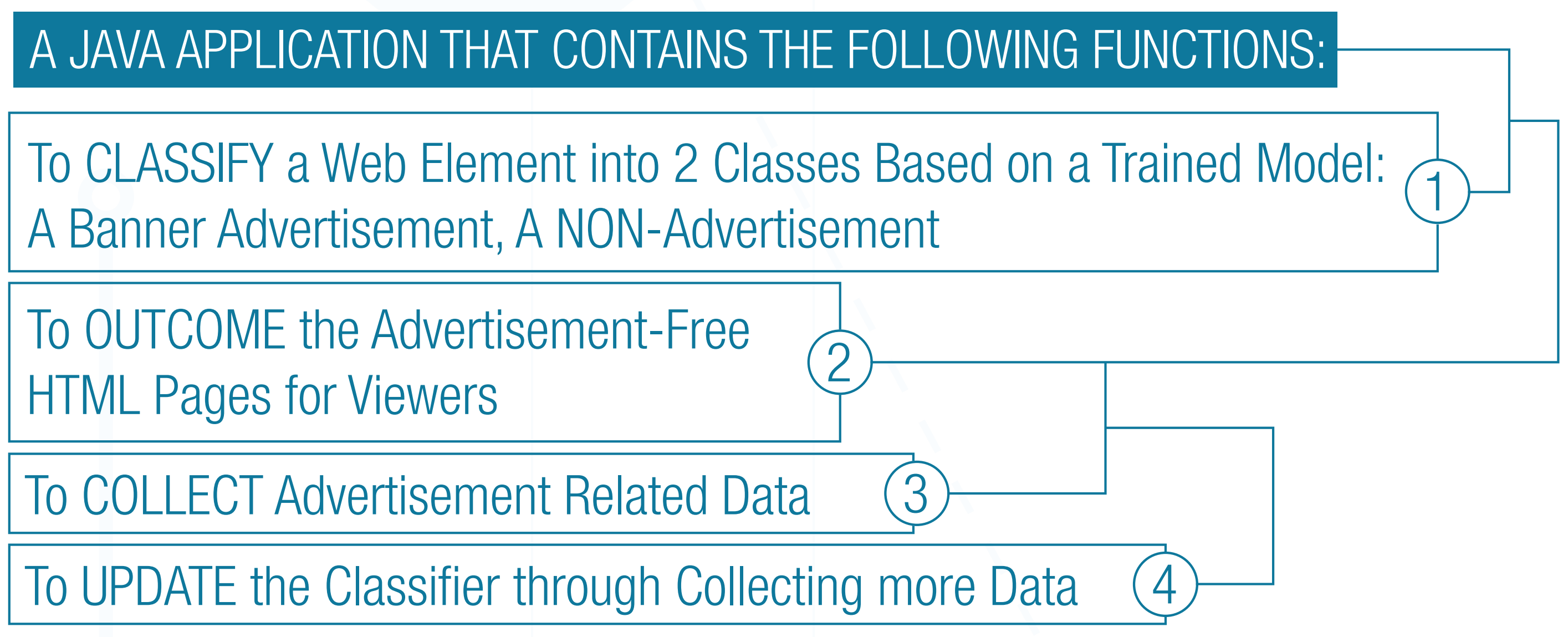
DEPARTMENT OF Electronic and Information Engineering 電子及資訊工程學系

The Hong Kong Polytechnic University
香港理工大學

DATA ANALYSIS IN THE INTERNET ADVERTISEMENTS

THIS PROJECT AIMS

to analyse web banner ad and detect if a website element is a banner advertisement or not. This is achieved through applying current data mining techniques such as data processing method, and classification algorithm. Solely web banner classification is considered because it is the primary advertisement on a web page and it is the easiest to obtain data. If the idea work on web advertisement detection, the same idea can be applied to other applications.



A JAVA APPLICATION THAT CONTAINS THE FOLLOWING FUNCTIONS:

1. To CLASSIFY a Web Element into 2 Classes Based on a Trained Model: A Banner Advertisement, A NON-Advertisement
2. To OUTCOME the Advertisement-Free HTML Pages for Viewers
3. To COLLECT Advertisement Related Data
4. To UPDATE the Classifier through Collecting more Data

CLASSIFICATION ALGORITHM SELECTION

Naïve Bayes Classification:
Naïve Bayes is considered because it is quite popular in text classification.

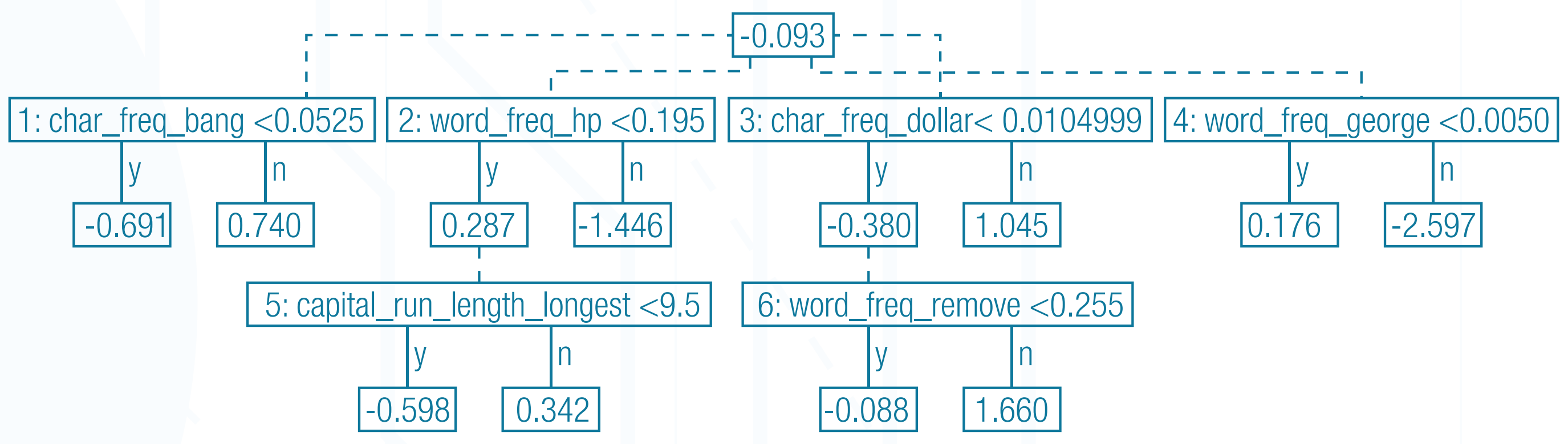
$$P(y|x1,x2,..,xn) = \frac{P(y)\prod_i^n P(x|iy)}{P(x1,x2,..,xn)}$$

y : class y x : features

There are number of reasons to choose consider Naïve Bayes Classification. Firstly, It can handle missing values in features. It is also effective when the dataset is limited with its high bias ,low variance characteristics.

CLASSIFICATION ALGORITHM SELECTION CONT

ADtree (Alternating decision tree) is another potential candidate. It is an evolved version of decision tree with boosting technique. It consists of decision node ,used for predicate decision and prediction node , containing a single number . A class is determined by the summation of all prediction node score that it passed through. Figure(1) shows an example. It has been chosen because it is easy to interpret and it has smaller tree size. It also provide a best tree structure by a number of iteration.


-0.093
1: char_freq_bang <0.0525
2: word_freq_hp <0.195
3: char_freq_dollar< 0.0104999
4: word_freq_george <0.0050
y
n
-0.691
0.740
0.287
-1.446
-0.380
1.045
0.176
-2.597
5: capital_run_length_longest <9.5
6: word_freq_remove <0.255
-0.598
0.342
-0.088
1.660

REAL APPLICAITION

A real application is built for utilizing the classifier. It is a simple browser where user can surf the internet. It also provides users to parse advertisements data into txt file. It can generate or update current dataset to train a new classifier. The most important function will be detecting the advertisements. Figure 3 shows the GUI of the application. Figure 4 shows the effect on a detected advertisement. It is just hidden but it still occupies the position.

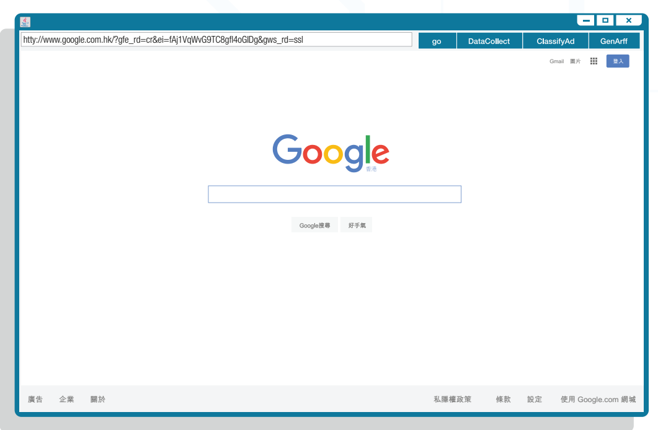

http://www.google.com.hk/?gfe_rd=cr&ei=fAj1VqWvG9TC8gfl4oGlDg&gws_rd=ssl
go
DataCollect
ClassifyAd
GenArff
Google

Figure 3 Application GUI

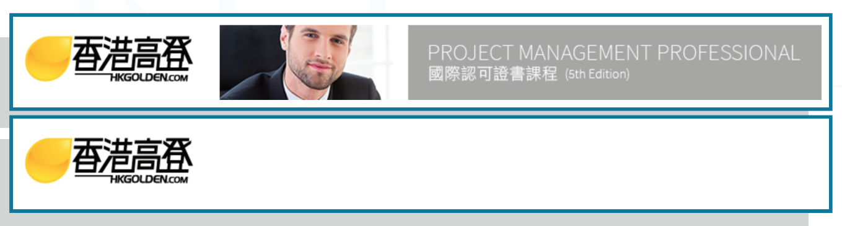
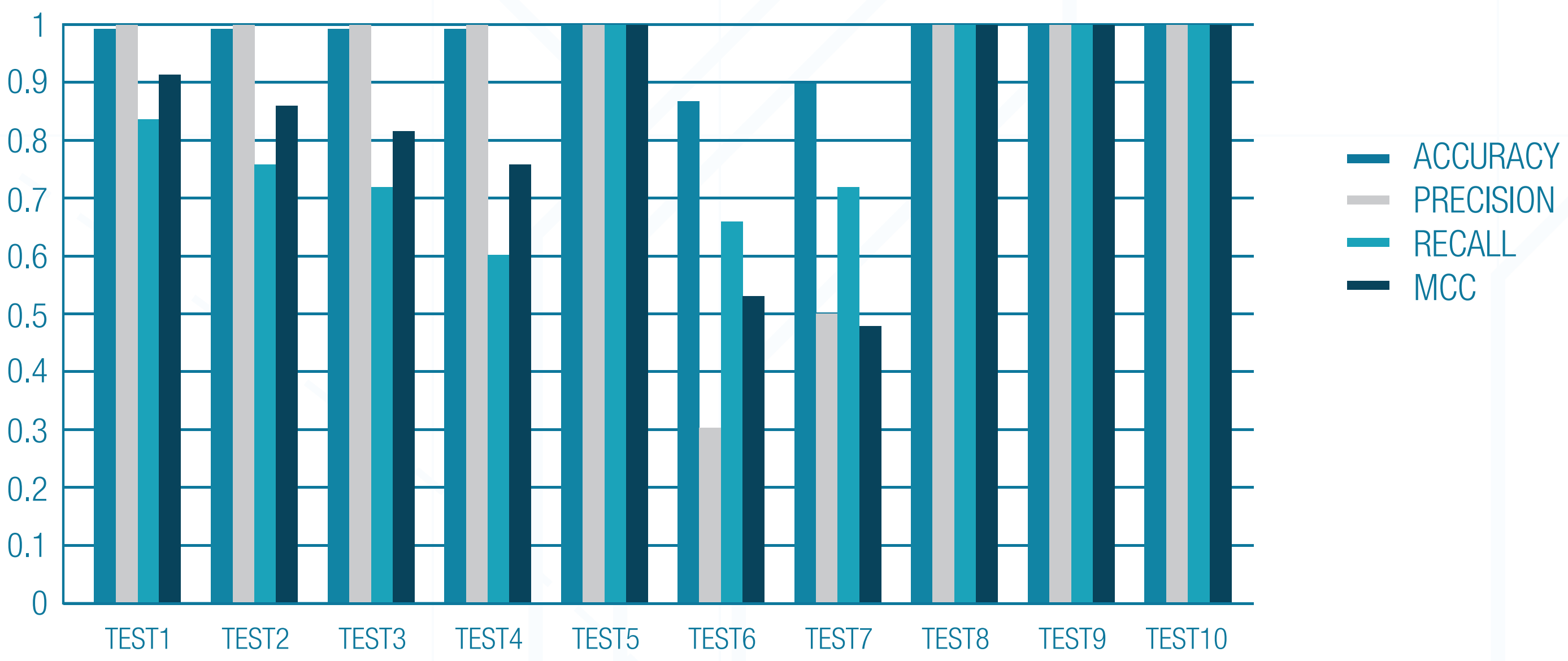

香港高登 HKGOLDEN.COM
PROJECT MANAGEMENT PROFESSIONAL
國際認可證書課程 (5th Edition)
香港高登 HKGOLDEN.COM

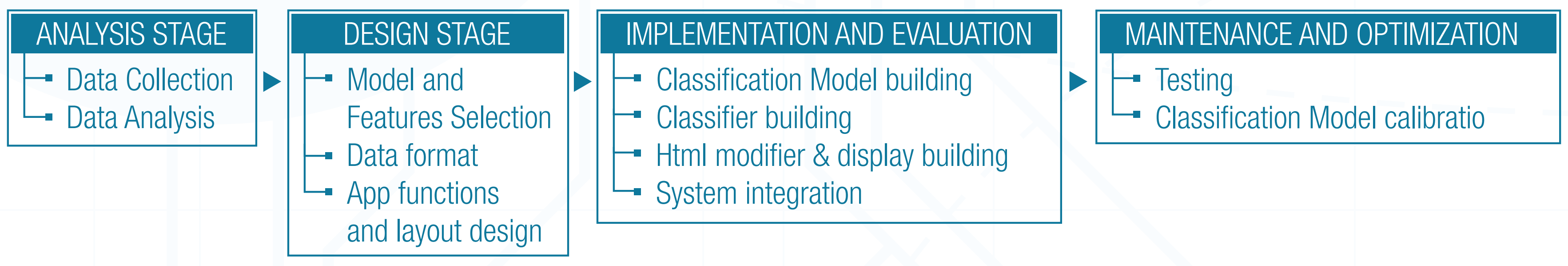
Figure 4 Effect of Detecting Advertisement

APPLICATION TESTING RESULT

10 tests on different websites have been conducted. They include Hkgolden.com, https://hk.yahoo.com/, Babykingdom.com, Sina.com, Pixnet.net, Cnn.com, Time.com, Gamespot.com, Expedia.com, Engadget.com. The result is shown in the below graphs:



METHODOLOGY



ANALYSIS STAGE
- Data Collection
- Data Analysis

DESIGN STAGE
- Model and Features Selection
- Data format
- App functions and layout design

IMPLEMENTATION AND EVALUATION
- Classification Model building
- Classifier building
- Html modifier & display building
- System integration

MAINTENANCE AND OPTIMIZATION
- Testing
- Classification Model calibratio

DATA COLLECTION AND DATA ANALYSIS

JxBrowser is a java library to the project to be the parsing tool. It is a library that embeds chromium-based Swing or JavaFX component into the Java application.. It also offers clear API documentation and community support so that problem can be solved promptly. Thankfully, with the help from Dr.Bonnie Law, the project is successfully endorsed as an academic project which can freely use the library in the FYP period.

Iframe and anchor tag with image inside are targeted. For a tag, half, image source, title, weight and Height, these attributes are obtained. For Iframe, ID, source and W&H are collected

After collecting around 400 instances of banner advertisement and 100 cases of non-banner advertisement, some features of advertisement can be found out.

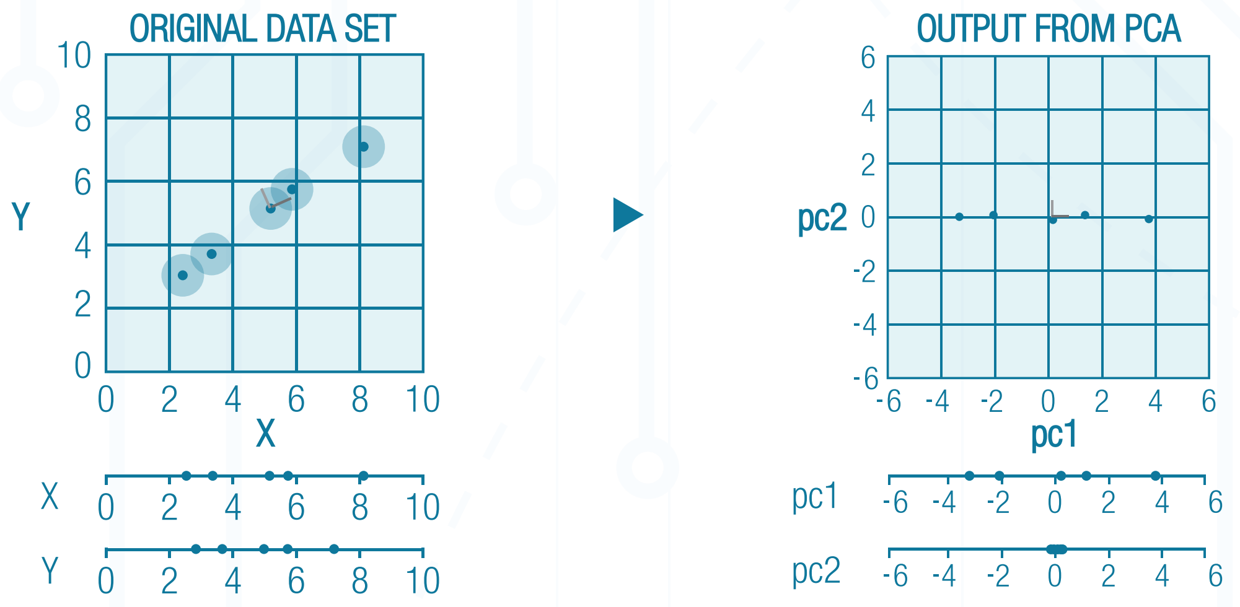
COMMON ADVERTISEMENTS DIMENSION				
300X250	728X90	180X150	250X50	300X50

Another finding is that around 85% of web banner advertisement is posted in Iframe format rather than image, and a possible reason is it is easier for the client to customize their advertisement since Iframe can load the different resource.

FEATURE SELECTION

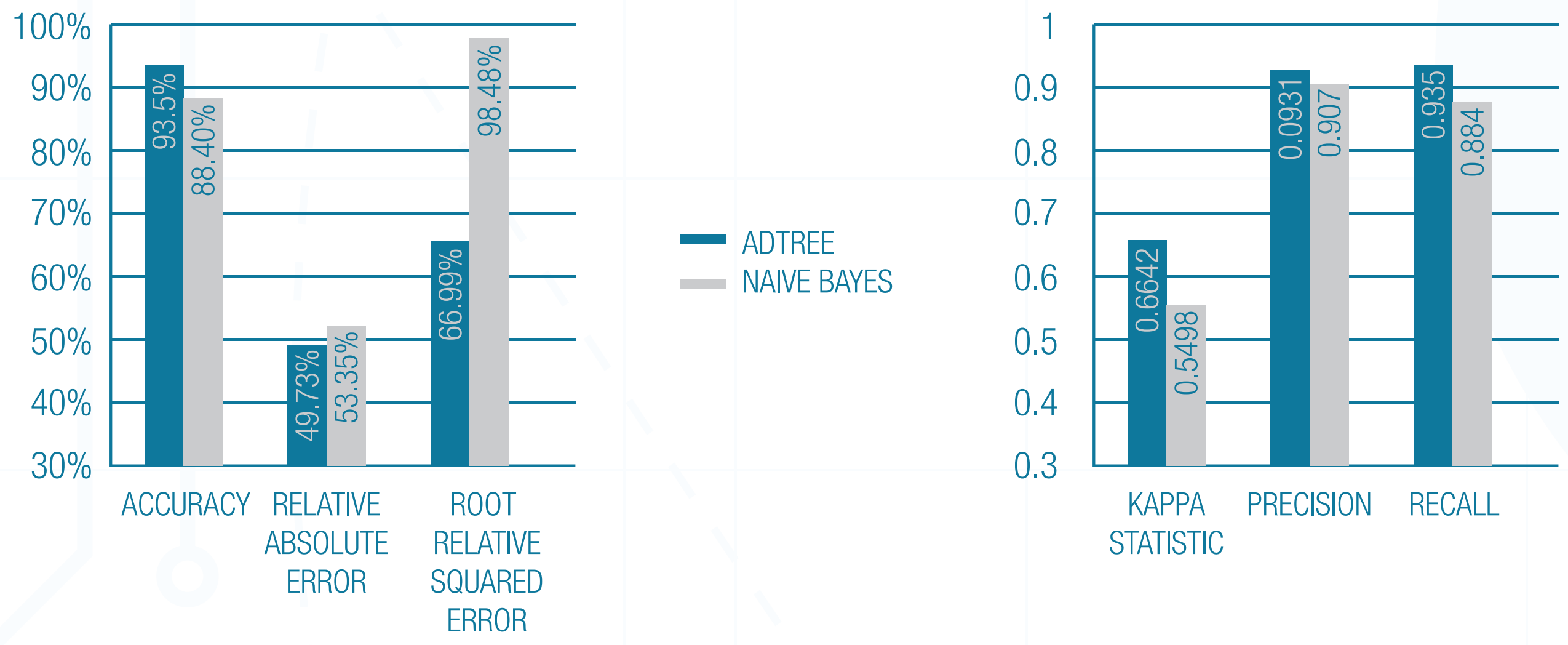
The feature selection is based on two criteria: quantity and significance of a word. The amount of a word means the number of occurrence in the data. If the frequency is high, the word will be selected. On the other hand, if a word can determine an instance if it is an advertisement easily, its significance is high, and it will be chosen though the quantity may not be significant.

Principle Component Analysis
Its main purpose is to attempt to find a group of linearly uncorrelated features/variables to interpret most information in the raw data. It first applies orthogonal transformation to data so that data are converted into a new coordinate system. The axes in the new coordinate system are actually the engine vectors of the data, which is the principle components. Figure 2 illustrates an example. It can be concluded that pc2 can be disposed after the transformation and principle component calculation. Since the variance of pc2 is zero so it is no use in classifying an instance. With the aid of PCA, dimension can be reduced and features can be more uncorrelated.



CLASSIFIERS EVALUATION

A statistical way of evaluation is implemented first before applying it to the real scenario. Ten-fold Cross Validation is used to test the dataset with1215 instances



ADtree is winner since it is better in every aspect. Therefore, ADtree will be the classifier in the real application

CONCLUSION

Machine learning is a viable way to detect advertisements. It doesn't need too much dataset to obtain a applicable classifier. It is also found out that the performance does not improve as the training sample increase. Therefore, more features should be introduced or a more advanced method could be used such as Deep Learning method.
Another concern is the speed problem. The average elapsed time on 10 test is 8.9 second. Generally speaking, the elapsed of loading a normal webpage is within 2 seconds so this implies the application need to improve the processing speed greatly.

Chan Chi Lam 12061255D   Supervisor(s): Dr. Bonnie Law