Students' Placement Office, IIT Kanpur

Title of the Project	Content Monitoring
Commencement Date	15-05-2013
Completion Date	19-07-2013
Project Supervisor	Dinesh Ajmera
Organization/Institution where the Project was accomplished	Walmartlabs, WM Global Technology Services India Pvt. Ltd, Bangalore

Project Description (You can use extra A4 sheets in case you run out of space however the extra sheets should also have the seal & signature of the Project Supervisor or the relevant authority)

Walmart Display its advertisement on various sites. But it doesn't want to show its ad on bad pages which contain Adult, Racist, Communal, etc content. For this it need to monitor the pages on which it could Display its ad. Overall project was to built a content monitoring module which could classify pages into two categories :

1. one which adher to walmart policies and we could display our ad on those pages
2. other which Doesn't adher to walmart policies and we shouldn't Display our ads on it

We were already using a classifer but were not sure about the performance. First part was to build a new project and implement the current classifier and then do a performance analysis. Current system had two parts :

1. Adult Word Percentage analysis: It calculates the percentage of adult words in the web page content and if it is greater than some threshold then classify page as Bad page otherwise good.
2. Alchemy Concept Analysis : Alchemy API is a Natural Language Processing service which provide different sematic analysis of data. We used concept tagging in which alchemy attached some concepts to the page and confidence(0-1) of that concept for that page.

We would like to have a system which has :

1. High Recall for Bad pages
2. High precision

Current System had poor performance with recall for bad page as 48.54% and precision as 75.50%
I built a Data Set and analysed different Threshold for both Adult Word Percentage analysis and Alchemy Cocept Analysis and increased the performance to recall for bad pages 76.70 % and precision 94.24%.
I furthur tried to improve the performance by using these features:

1. Word probability score : I generated the probability of occurrence of word in Bad page(P_b) and good page(P_g) by training on the content of 20,000 Bad pages and 20,000 Good Pages. This training set was built from 3 lakh Bad url and 2.5 lakh Good URLs after filtering URLs based on some criteria Then a score was assigned to each pages based on these probability.
2. Concept vs. concept Matrix score : Generated Concept vs concept Matrix to capture the context in which concept is occuring. Again Matrix was generated on the above mentioned Dataset.
3. Some other features like Bad concept confidence sum, Bad concept count, etc were also introduced.

Name : Dinesh Ajmera
Designation : Director
Signature: _

I generated these features score for the data set and ran logistic regression and OneR rule to Train model on some training set and then put the parameter generated in the actual code. I tried different combinations of feature to remove noise and get the best results. Performance increased to

1. Recall for bad pages – 93%
2. Precision – 96%

I also implemented stemming of words using porter stemmer. But final code doesn't include it. Also I looked into classification based only on URLs. This doesn't include domain level classifcation but we looked at the structure of URL and tried to classify . Performance was quite good(recall around 70%). But all the URL caught on URL based classifcation were also caught by our new system. So that was removed from final Code.

Project Verification Form

By appending your signatures to this form you acknowledge and agree that:

- This form along with the certificate would serve as the official document between the project supervisor and Students Placement Office, IIT Kanpur regarding verification of the student's project work
- The student will provide additional information and documentation relevant to his/her project upon request by the Students' Placement Office
- The student has clearly defined his/her individual role in projects done in cooperation with other students, faculty, groups or company personnel.
- Incorrectly over-stating the reach, impact and/or quantitative/qualitative results of a project is unethical.
- In case of violation of any of the above rules, Students' Placement Office, IIT Kanpur reserves the right to take necessary action including de-registering the student from the placement season and reporting the misconduct to the Institute Authorities.

Submitted by:-	**Project Supervisor Details:-**
Name:Nikhil Aggarwal	Name: Dinesh Ajmera
Roll No: 10446	Designation:Director
Signature:	Signature: