METU, Department Of Computer Engineering

Graduation Project
Proposal Form

Project Information

Title

Web site crawler and categorization engine

Target

Public [X] Restricted [ ]

Proposer Information

Name(s)	*Alptugay DEĞİRMENCİOĞLU, LABRIS NETWORKS*
E-Mail(s)	*alptugay@labrisnetworks.com*

IP (Intellectual Property) Information

This is a public project which is expected to be released in BSD license.

Project Description and Background Information

Description

The project will have a crawler which will crawl the web and after the websites will be categorized (chat, adult, e-commerce, blog etc.) using machine learning techniques. The crawled results will be stored in a database. The system should be queried via an API. If the website queried is not in the database, it should be analyzed immediately and the category of the website should be returned to the user.

Similar Products/Projects

Zvelo [1]
Cyberoam [2]

Justification of the proposal

There are some open source web crawlers and closed source categorization engines. But there isn't any open source categorization engines and none of the products target Turkish websites. So this project will be the first website categorization project that will be open source and target Turkish websites.

Contributions, Innovation and Originality Aspects of the Project

There are some open source web crawlers and closed source categorization engines. But there isn't any open source categorization engines and none of the products target Turkish websites. So this project will be the first website categorization project that will be open source and target Turkish websites.

Technical Aspects of the Project

There are some open source web crawlers and closed source categorization engines. But there isn't any open source categorization engines and none of the products target Turkish websites. So this project will be the first website categorization project that will be open source and target Turkish websites.

Targeted Output, Targeted User/Domain Profile

The targeted users are all companies using web filtering products.
If the URL is already categorized and in database you should return an immediate response, if the URL is uncategorized you should categorize it and return a response in 15 seconds.

Project Development Environment

The project should be developed on an rpm based linux enviroment preferably CentOS.
Any programming languages and machine learning libraries can be used. For example weka or pyBrain.
Scrum will be used as the development method.
Git should be used as the version control system.

External Support

An advisor will be provided by the company to interested groups.

References

[1] https://zvelo.com/
[2] http://www.cyberoamsecuritycenter.com/cyberoamsupport/webpages/webcat/webcathome.jsp