Sprint Evaluation

Tasks:
- Database Connection (JDBC)
- Database Manipulation Functions
  - addURL, deleteURL, getCategory...
- Crawler Setup and Trials
  - Crawler4j
  - Webspinx
  - Scrapy
- Document Representations
  - Bag of Words
- Machine Learning Algorithms
  - WEKA
  - Attribute Relation File Format (ARFF) Research
- Information gathered on
  - Linear Regression with One Variable
  - Linear Regression with Multiple Variables
  - Logistic Regression

Achieved Goals:
- Database is working and Functions related to DB manipulation required for project are also ready.
- Web Crawler is also working and gathering Web Page data as txt file.

Overcome Problems:
- Web Page data gathering (HTML to TXT) is not a problem anymore since the crawler we are using now is capable of saving data as txt file.
- Different coding languages is not a problem anymore. We transfered all code segments to java language.

Plan Update:
- We need a translator program for arff file creation. (TXT to ARFF)

Team evaluation

Our team can work together just fine. Every week before our meetings with our assistant we meet and share what we researched or developed. Whenever we need a task to be done for the project we assign a team member for the project. The component-Team member associations are nearly finished meaning every member has a component that he/she wants to work on.

Finished Tasks:

Task	Assigned Member	1st week	2nd week	3rd week
Database Connection	Onur Ozan Yüksel	√		
WEKA Analysis	İzzet Barış Öztürk	√		
Web Crawler Research/Identifying	Mert Basmacı, Özge Donmaz	√		
Bag of Words Research	İzzet Barış Öztürk		√	
Crawler4J Install and Use	Mert Basmacı		√	
Regression Research	Özge Donmaz		√	
ARFF Format Understanding	İzzet Barış Öztürk, Özge Donmaz		√	√
DB Manipulation Functions	Onur Ozan Yüksel		√	√
Data Gathering from Crawled Web Pages	Mert Basmacı			√

Sprint II Backlog Updates

We had 4 Tasks for second sprint in the Start-up Document. 2 of them (3rd&4th) was decided to be finished on this sprint and 2 of them (5th&6th) was decided to be started in the 2nd sprint and finished in 3rd sprint. We managed to finish 2 of them but Data Sample gathering task couldn't be finished since we decided to finish Crawler task first and use the Crawler to get Data Samples. Instead of finishing 4th we finished 5th first. So there is a change in plans but we are still on track:

Task Number	Name	Description
T3	Database	Creating a database that meets requirements in a desired environment.
T4	Samples	Creating enough sample web pages to train the Machine Learning Algorithm.
T5	Crawler	A crawler to get data for categorization process.
T6	Training	Training the categorizing program via Machine Learning.

We finished 3rd and 5th tasks and in next sprint we will use the Crawler to get data samples and use them to train Machine Learning algorithm. The 4th & 6th Tasks will be finished in 3rd sprint. The rest of the tasks will be handled in 3rd sprint as planned.