Sprint Evaluation

Tasks:

- Text conversion with WEKA (research)
- Bag of Words Implementation (with static and dynamic bag)
- Turkish Language Detection
- Data Set Creation
- Getting content of a URL without crawler class
- Editing ARFF
- Turkish word list
- Base64 Encode/Decode
- WebCat Web Page
- MySQL to PostgreSQL

Achieved Goals:

- We managed to convert txt data into a bag of word vector that can be used for training Machine Learning.
- We succesfully got our initial training data sets for a few categories
- Started trying differrent Machine Learning Algorithms using the dataset

Overcome Problems:

- Due to the result of our requirements meeting with Labris, we had to change our Database type from MySQL to PostgreSQL. We successfully adapted to the requirement change.
- Our Crawler used to get all kinds of webpages with different languages. To create a categorizer training dataset, we eliminated all but Turkish web pages using a Language Detection Library.

Plan Update:

- Database Server Type Change
- Decided to save URLs as Base64 converted version
- Decided to make a different component than crawler to get data from a web page if not found on database.

Team evaluation

Now our roles in the team got determined. Onur is mainly responsible for Database Storage and Interconnection of the components of the project. Özge and Barış are responsible of Machine Learning Algorithm Training and Categorizer Component, and Mert is responsible of Crawler Component and Data Set Creation. Overcode team works as a team so if a problem occurs on any part of the project we all think of a solution and overcome the problem. Our team synchronization is really good. We'll finish the project according to the plan.

Finished Tasks:

Task	Assigned Member	1st week	2nd week	3rd week
Text conversion with WEKA (research)	Özge Donmaz	√		
Data Set Creation	Mert Basmacı	√		
Turkish word list	İzzet Barış Öztürk	√		
Base64 Encode/Decode	Onur Ozan Yüksel	√		
Bag of Words Implementation(Dynamic)	Özge Donmaz		√	
WebCat Web Page	Onur Ozan Yüksel		√	√
Editing ARFF	İzzet Barış Öztürk		√	√
Bag of Words Implementation(Static)	Özge Donmaz			√
Turkish Language Detection	Mert Basmacı			√
Getting content of a URL without crawler class	Mert Basmacı			√
MySQL to PostgreSQL	Onur Ozan Yüksel			√

Sprint III Backlog Updates

We had 5 tasks (5-9) for third sprint in the Start-Up Document. We already finished 5th (Crawler) task in 2nd sprint but couldn't finish 4th (Samples) task due to a change of plans. So in this sprint we finished creating sample data sets for training(T4), training the categorizing program with it (T6), and creating a categorizing program (T7). After the meeting with Labris, we decided that our project won't need an Interface since it will be called by an API. So we won't be doing 8th Task.

Task Number	Name	Description
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
T7	Categorizer	A fully working categorizing program that uses data that uses crawler
T8	Interface	An interface for final user usage
T9	Optimization	Crawling and Categorization optimization that enables the project to meet the 15 second limitation.

We couldn't yet created a finished product to optimize so couldn't do 9th Task. But it was a miscalculation of our planing, we were supposed to do this task at the end of second midterm. Thus we accomplished this semesters goals and finished all the tasks. We'll try to make more careful estimations next semester.