Retrospective Document

Sprint 2

Work & Test Progress

Milestone 1: *Training Set for 50 Categories Expanding the training datasets to cover 50 categories.*

With 100% we have completed this milestone. We have collected 30 URLs for each category. In order to get supervised data, we have chosen URLs by manually searching that category name in Google. Then we choose more proper web sites which contains the most related textual content. Before we train data, we shuffle them, then train 80% of it. Then we use %20 of it in order to test accuracy of our different classifiers.

Milestone 2: *Training Set for 100 Categories Expanding the training datasets to cover 100 categories.*

With 100% we have completed this milestone. We have collected 30 URLs for each category. In order to get supervised data, we have chosen URLs by manually searching that category name in Google. Then we choose more proper web sites which contains the most related textual content. Before we train data, we shuffle them, then train 80% of it. Then we use %20 of it in order to test accuracy of our different classifiers.

Milestone 3: *Training Set for ALL Categories We will try to cover the rest of the categories that our final product is required to. (Expanding the training datasets to cover 141 categories)*

Because of the indistinguishable categories, insufficient samples, it is not possible for us to collect 141 categories. Moreover, some categories are very similar and we want to get rid of some of them. It is must, because it also increases classifiers' accuracy.

Milestone 4: *Categorization Algorithms Trying out diferrent algorithms and finding the best categorization algorithm possibble for our project.*

With 80% we have completed this milestone. It is not 100% since there is no limitation to try a new classification algorithm. We have also arranged a meeting with Ayşenur Birtürk Hoca to discover new classification algorithms for Turkish Language. Until now, we have tested 3 different classifiers and we are trying to implement a classifier for doc2vec feature vector.

TESTS:

1. *Shuffling the collected data and splitting into 2 with (4:1) ratio for training and test purposes.*
   - *Demo will contain a smaller version of the datasets but we'll show whole collected data.*
2. *The system that processes the collected data and creates a training and a test arff file out of them.*
   - *Demo will show the created arff files again with smaller dataset.*
3. *Model creation with the same training arff but different algorithms.*
   - *Demo will show 3 different trained model and a sample testing for each of them.*
4. *Statistical information of trained models using test arff*
   - *We'll show how well the trained models work for test cases.*
5. *Classification with logical regression classifier which uses doc2vec traning vector.*
   - *Note that this classification method may not be proper for our future purpose. Although it is not proper, it may be adapted to WEKA or other classification algorithms may be applied on doc2vec vector. Moreover, we have implemented it in order to investigate other categorization algorithms. Therefore we will show it on demo.*

Team Progress

- *Onur Ozan Yüksel 25%*
- *İzzet Barış Öztürk 25%*
- *Mert Basmacı 25%*
- *Özge Donmaz 25%*

Left-overs (Backlog)

Due to various reasons some of the categories' data samples couldn't be completed thus Milestone 3 is not 100% completed. Some of the reasons were: - Insufficent samples on internet due to category being too spesific - 2 categories being too similar - Since the scope is only turkish webpages we need to find web pages in a spesific language and some of the categories have little Turkish samples.

Next Sprint

Milestone 1: *Optimizing the ML algorithms changing the parameters of the algorithms.*

Milestone 2: *Exception handling throughout the system(from GUI input to category input).*

Milestone 3: *Expanding data samples of each category.*

Milestone 4: *Trying new classifiers and new categorization algorithms.*

Assistant's Evaluation

Assistant's (Team Leader's) comments regarding to this completed sprint.

Supervisors's Evaluation

Supervisor's (Team Leader's) comments regarding to this completed sprint.