Work & Test Progress

**Implementation Milestones:**
- **Doc2Vec Implementation** %100
  Deep learning via the distributed memory and distributed bag of words models.
- **Zemberek Root Implementation** %100
  Optimizing vector creation ML. (both training and classifying)
- **URL to ARFF (end-to-end) Implementation** %100
  Input: URL list, Output: Training + Testing ARFF files.
- **8 Different ML Algorithm Implementation** %100
  Algorithms: J48, ZeroR, ConjunctiveRule, OneR, DecisionStump, RandomTree, FilteredClassifier, Dagging

**Testing Milestones:**
- **3 category classifying test** %100
- **51 category classifying test** %100
- **76 category classifying test** %100
- **Doc2Vec classification using infer vector** %100
  A shorter vector used for better time/precision.

Team Progress

- **Onur Ozan Yüksel 20%**
- **İzzet Barış Öztürk 30%**
- **Mert Basmacı 30%**
- **Özge Donmaz 20%**
Left-overs (Backlog)

- Sample size for each category could not be increased due to time constraints and other problems that occurred while testing.
- Zemberek Rooted version couldn't be tested since it takes too long to create vectors with that method.

Next Sprint

**Imp. Milestone 1) Optimizing the ML algorithms changing the parameters of the algorithms.**
- Increasing classification accuracy

**Imp. Milestone 2) Crawler - Classifier - Database (The whole system) integration**
- Finalizing the project

**Testing Milestone 1) Weekly Testing 1**
- Testing results of optimization changes

**Testing Milestone 2) Weekly Testing 2**
- Testing results of optimization changes

**Testing Milestone 3) Weekly Testing 3**
- Testing results of optimization changes

**Testing Milestone 4) Weekly Testing 4**
- Testing results of optimization changes
Team Comments

The team got side-tracked because of the new approaches that are recommended by supervisors and other teachers. Also even if root finding is really fast when done sequentially it works really slow (1 webpage = 2 min.). We should use threaded approach for that.
Supervisor Evaluation