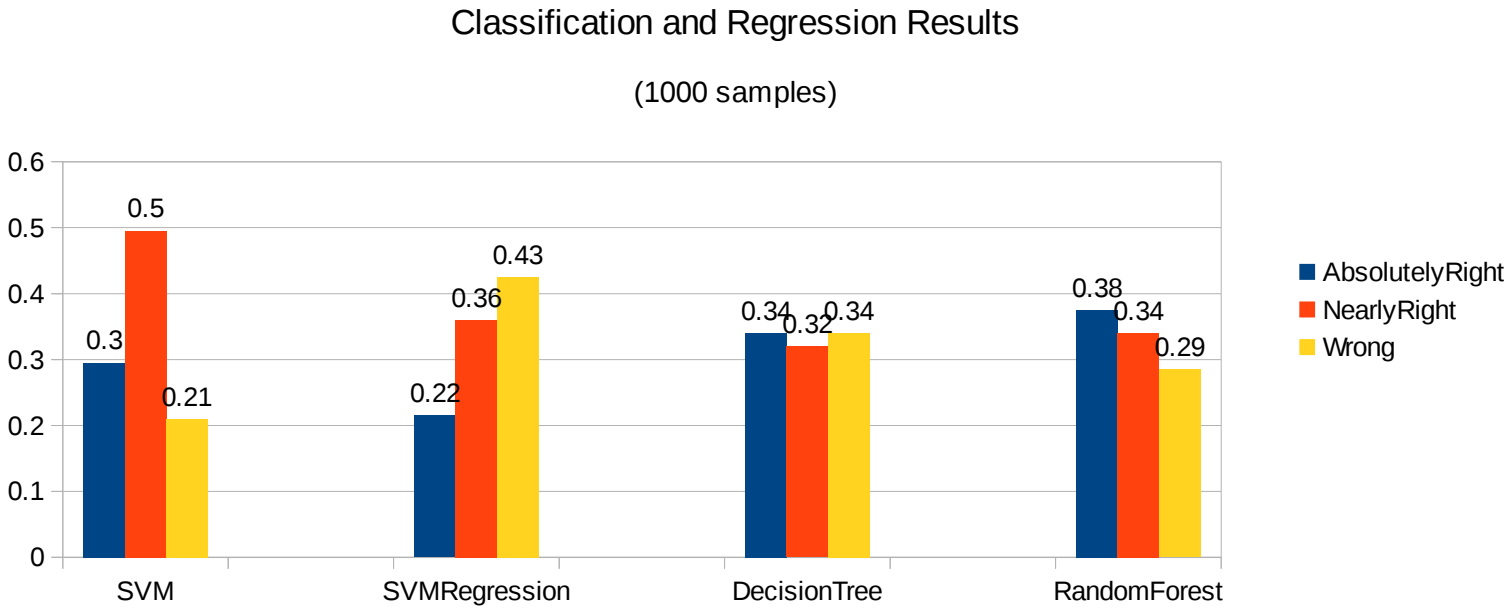
YELP-DS SEMESTER PROJECT
WEEKLY REPORT

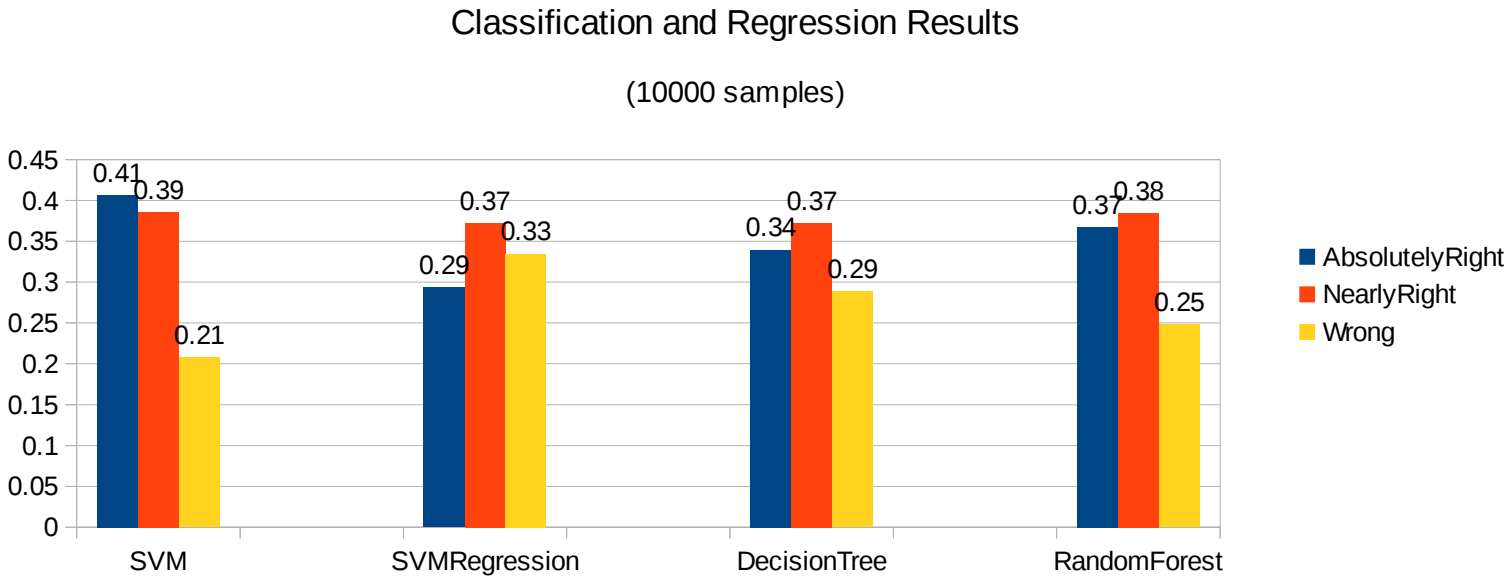
I. Objectives

1) Randomly select (Cross validation) some reviews as the training and testing dataset.
2) Try SVM with Scikit (linear kernel) to improve the speed when training the model.
3) Try to use Decision tree (Random forest) or Neuron networks as the classifiers.
4) Read the book - An introduction to statistical learning.
5) Try regression instead of classification methods.
6) Try to give graphical results.
7) Future work: use spark for parallel computation and combine multiple models for higher accuracy.

II. Details

#TASK	STATUS	DETAILS
1	Done	Randomly select an amount of input data to build the model (training data ~ 80%) and test (testing data ~ 20%), because we need to install the Scikit library to run the code so we couldn't run it on the Lab's machines. That is the reason why we've just tried to run with 1000 and 10000 samples to compare the results of multiple models in different number of data.
2	Done	We've already tried Scikit library with SVM with different kernels (linear, rbf, polynomial) and recognize that the linear one is fastest one and the polynomial is the most accuracy one. But the polynomial is really time consuming while the result of the linear is acceptable, so we choose the linear kernel as the candidate for SVM.
3	**In progress**	We also tried the Decision Tree and Random Forest as classifiers. The results are not bad but they are not as good as the SVM case. **Now we're trying to understand the reason.**
4	Done	Because the time is limited so we just read the overview of the book to understand the basic ideas and the notions in order to do our job.
5	**In progress**	With the regression one, we got the result but we don't know how to evaluate it. First, we rounded the results and compare to the original values. And it's not very good. **May be we can ask you more about this.**
6	Done	We made some charts based on the results we got so that you can see and compare (in the figures below).
7	**In progress**	Next week may be we can **try to use SPARK (improve the speed)** to do the experiment with different parameters on the whole dataset. Because we think it's better to test it with all data (not part of data) to see which parameters are good. We think that's a problem because we can't run it on the lab's machines (lack of libraries). Along with that, we also try to **use bagging classifier (combining multiple models)** to **improve the accuracy**. Now the accuracy (optimistic) is ~ 80%, our objective is to make it increase to 90%.





Explaination:
- AbsolutelyRight: “predict label == original label”
- NearlyRight: “|predict label - original label| <= 1”
- Wrong: the rest cases.

Notes: We can show you the details about the results (about the visualization of the decision tree and the evaluations for every model) when we meet.