Homework 3

1. Downloaded the test and train dataset successfully i.e. qb.train.csv and qb.test.csv.

2.
   a. I tried to create a classifier using the training data and I used SVM, decision tree and logistic regression one by one to check the accuracy of the model in each of the method. I have performed the following steps to create the classification-
      - Uploaded the dataset in Python.
      - Defined the matrix of features(x) using relevant columns and dependent vector(y).
      - Preprocessed the data by applying LabelEncoder on the categorical columns.
      - Imported re, nltk to create corpus of the text containing columns.
      - Created dummy variables for columns using OneHotEncoder.
      - Imported train_test_split from sklearn to apply cross validation on data by splitting the data into training and test set.
      - Performed feature scaling using StandardScaler from sklearn.
      - Imported LogisticRegression from sklearn to build a model.
      - Imported DecisionTreeClassifier from sklearn to build a decision tree model.
      - Imported SVC from sklearn to build an SVM model using 'rbf' kernel.
      - Imported confusion_matrix from sklearn to build an error metrics for all the models.
      - Imported accuracy_score from sklearn to find the accuracy of the models.

Features	Accuracy		
	Decision Tree	SVM	Logistic Regression
All variables	57.54	46.53	54.26
Excluding text, page, answer	72.77	66.77	66.7

   b. I found out the accuracy of each model by including all the variables. On checking the accuracy by considering different combinations of variables, I found out that keeping the variables text, page and answer together or individually in the model was limiting the accuracy of the model. Here, the pattern could be seen that the presence of each of these 3 variables either individually or together made no sense to the model as it was keeping the accuracy of the model same with no improvement. This pattern was seen in all the models and I decided to remove these three variables. Hence, on removing all the three variables I got increased accuracy as shown in the table. The decision tree model gave the highest accuracy of 72.77%.

3.

a. In this part, I tried to find out which parameters are more influential in defining the accuracy of the model. I found out that the length of the text and the year in which tournament was held had an impact on the accuracy of the model. I have made 2 different plots: length vs answer and tournament_year vs answer which helps to understand the relation between the parameter and the chances of an answer being correct. Looking at the plots, it can be seen that the chances of an answer being correct or incorrect are almost equal for all tournament years except years 2005, 2007, 2008 and 2009. Also, if we look at the length of the text, we can see that most of text falls in the range of 10 to 150 in terms of length which is the main corpus for observation. Thus, I have included two new columns in the dataset for length and tournament_year so as to improve the performance of the model. Also, I have analyzed the that if the answer_type is "work" there are more chances for the answer to be correct.

   I have performed the following steps to create the improve the model-

   - Uploaded the dataset in Python.
   - Imported re and created a new column as tournament_year by separating the year from the existing 'tournament' column.
   - Created a new column to calculate the length of the text in the column 'Text'.
   - Defined the matrix of features(x) using relevant columns and dependent vector(y).
   - Imported re, nltk to create corpus of the text containing columns.
   - Created dummy variables for columns using OneHotEncoder.
   - Imported train_test_split from sklearn to apply cross validation on data by splitting the data into training and test set.
   - Performed feature scaling using StandardScaler from sklearn.
   - Imported LogisticRegression from sklearn to build a model.
   - Imported DecisionTreeClassifier from sklearn to build a decision tree model.
   - Imported SVC from sklearn to build an SVM model using 'rbf' kernel.
   - Imported confusion_matrix from sklearn to build an error metrics for all the models.
   - Imported accuracy_score from sklearn to find the accuracy of the models.

b. On adding the new parameters to build the model, I got an improved accuracy in the models as follows-

Features	Accuracy		
	Decision Tree	SVM	Logistic Regression
Include Length and tournament year Exclude page, text and answer	76.73	68.87	68.31

As we can see here, there has been an improvement of around 2% in SVM and Logistic Regression models and an improvement of around 4% in the decision tree model. Hence, I will be choosing the decision tree model to make predictions on the given test data.