Machine Learning Project
Team Member: Shirong Bai, Xinshuo Yang and Xiaohang Zhang

Project: We are going to work on the default project, which is answering questions about science.
How to get data: Use Illinois Wikifier code to extract useful data from Wikipedia, if time allows, we might try to extract data from other scientific websites, such as Chemwiki, Biowiki, Physwiki, Geowiki, Statwiki, Mathwiki, Solarwiki, Wikibooks, Math.Wikia and WikiTrivia, etc. Or forums like Quora, Answers, Wolframalpha, and Ask, etc.

Techniques:

1. Feature engineering
    a. Stemming/lemmatization using NLTK, remove stop words.
    b. N-grams
    c. Use sklearn.feature_extraction.text.CountVectorizer to convert text document to token matrix or use TfidfVectorizer.
    d. Group text into different categories like Math, Physics or Chemistry, etc.
    e. Sentiment analysis: Using the Pattern.en library.
    f. Identify proper verbs, nouns and adjective, etc. Probably, Use Parts of speech .(POS) as features and assign them weight in the end.
    g. Sentence modality: Use the modality() function from the Pattern.en library
    h. Select significant features, filter out unimportant feature. For example, we can play with sklearn.feature_selection.SelectKBest to select the features of interest to us in this project.
    i. An question we want to ask is how to select labels, obviously this is not a trivial classification problem we used to see in homework, we got to decide what labels we want to use. Our idea at this time, is, to extract potential labels from choices.
2. Training techniques
    a. Logistic regression construct a feature vector and label vector for further operation
    b. Linear regression
    c. Naive bayesian classification
    d. SVM
    e. Perceptron
    f. Decision Tree
    g. Boosting

Timeline:

1. November 13th, finish basic framework, should have a output file "sci_sample.csv" with correct format.
2. December 1st, try all the techniques listed above, get a decent ranking hopefully.
3. December 15th, finish the final presentation.