OSDC/PIRE Research Proposal:
Predict the click-through rate of ads given the query and user information.

http://www.kddcup2012.org/c/kddcup2012-track2

The task is predicting the click through rate (CTR) of advertisement, meaning that we are to predict the probability of each ad being clicked.

Knowledge/Skill requirements:

Advertisement: click-through rate (CTR) = #clicks / #impressions

Statistics:
- ROC/AUC, MSE(mean squared error), RMSE(root mean squared error) Weighted mean,

Data-mining:
- Linear (and non-linear?) classifiers, Supervised Classification, Logistic Regression, Naive Bayes, SVM, Perceptron, PA(Passive Aggressive), EM algorithm, Ensemble Learning, AdaBoost.

IR:
- Matrix Factorization, Cosine Similarity, TF-IDF, Bag Of Keypoints (BoK), (Latent) Topic Model, LDA

Data processing:
- Hive, Pig, Hadoop, Mahout, Jubatus, R

GOAL

AUC higher than 0.8. The best score in the contest is \textbf{0.80891}.
The baseline is \textbf{0.71198} when using AD id for features (see basic_id_benchmark.py).

http://www.kddcup2012.org/c/kddcup2012-track2/leaderboard

The average CTR of training dataset is 0.0387.

Implementation Goal
Extending Pig or Hive (as their UDFs) for prediction.

Design choice: (1 is preferred)
1. Create custom pig scripts (or Hive UDFs) of linear classifiers.
2. Considerable to use Jubatus, Weka, libSVM, Mahout in the UDFs.

Possible Research Target
Try to write a paper if the AUC could get around 0.8 or more.

Detailed Info: