Click stream analysis

Team members:
Sai Santhosh Vaidyam Anandan (SBU ID: 109819365)
Sowmiya Narayan Srinath (SBU ID: 109747156)

Assumptions:
The input and training data files are present in the same directory as the code.
Following are the files used by the program:
testfeat.csv testlabs.csv trainfeat.csv trainlabs.csv

Logic and Execution:
Initially the program parses the training data as a numpy array. All the numerical values have been divided into three ranges: *high,med and low* so that the decision tree can be built in finite time with the branches at a manageable amount.
Since taking all 274 pages as features leads to extremely deep trees with long time for building and traversal , we only pick pages with more random distribution as features.
This is controlled by the *num_features* value in learning.py. As the number of features is reduced the runtime is reduced but this comes at the cost of loss of accuracy in prediction.

Auxiliary functions used:

build_test_data and build_data : To convert the input from training and test data to a numpy array readable by the other modules
calc_h_to_x : To calculate the entropy value H(X) for training data set
split_on : Calculate the information gain values IG(X) for each of the features by calculating h(x|low),h(x|med) and H(x|high). Then it wil return the column (feature) with highest entropy
search_tree : Recursively traverse the ID3 decision tree built by the learning phase and return the leaf value.

Output Trace:

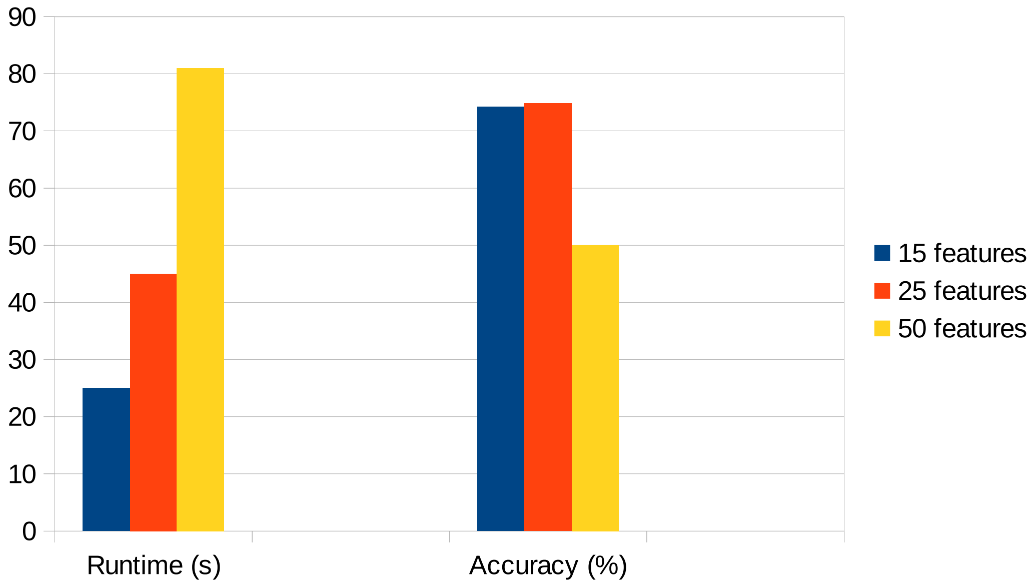
```
 $ python click_stream.py
Loading data ...
Data loaded
Building the decision tree, please wait...
Decision tree built!
74.836 % predictions are correct.
Please check results.txt for the predictions...
Runtime:  45.6355888844 s

 $ more results.txt
0
0
1
```

```
1
0
0
0
0
0
.
.
.
```

Results Analysis:

num_features	15	25	50
Runtime	**25s**	**45s**	**80s**
Accuracy	**74.23%**	**74.83%**	**78%**

The key parameters in comparing the number of features are the **runtime** and **accuracy of prediction**. Below graph compares the num_features on these parameters for given input data set.



Files Submitted:

click_stream.py and learning.py – source code for detection and learning phases
testfeat.csv testlabs.csv trainfeat.csv trainlabs.csv – input test and tranining data sets
results.txt – sample output for one run