User Classification Write Up

Background

At The Home Depot, customers can be broadly classified into two categories: professionals and non-professionals. Professionals consist primarily of contractors and those making purchases on their behalf while non-professionals are customers that are typically shopping for supplies for DIY home improvement projects. The two groups have different needs and require different levels of assistance. In the context of search, a professional will most likely have a better idea of specific products he or she is looking for when visiting our site, while a non-professional might have more of a general idea of the task they need to accomplish but not necessarily the products they will need to do so. Knowing which type of customer has made a query can help us to provide a better user experience by returning more personalized query results.

Problem Description

Your task is to design an algorithm that can determine whether a query was generated by a professional or a non-professional. Given a query your algorithm should output either 'Pro' or 'Non-Pro'.

Training Data

The training data will be provided in a text file called 'search_terms_train.txt'. Each line of the text will have a query string and a label of either 'Pro' or 'Non-Pro' in the following format-

```
search term|label
```

Testing

Input

The input for testing will be a text file called 'search_terms_test_no_label.txt'. Each line of the file will contain one query string to be classified.

Output

The output of your program will be a text file with the query strings from the testing input file and the labels generated by your team's algorithm. Each line should be in the same format as the lines in the training data file-

```
search term|label
```

The output file should list the query strings in the same order as they are listed in the input file.

Scoring

Scores will be determined by the F1 error of correctly labeling a query as 'Pro'