This file describes the Algorithm used by team – Aardvarks. The final model is an ensemble of eight individual models.

Software Used:

Platform: Windows (x64)

R: 2.13.0
randomForest package: 4.6-2

Python: 2.7.1
Numpy package: 1.6.0
Also add python.exe in the system path

How to Run:

1. Setup Directory Structure
   (a) Make a folder on your local machine and import the raw data files "training.tsv", "regdate.tsv", and "wikichallenge_example_entry.csv" in that folder. This folder is referred to as DATADIR in the code.

   (b) Extract the contents of the zip file on your local machine. The top level directory is referred to as MAIN_DIR in the code. MAIN_DIR should contain following files and folders upon extraction:
   - "AllFeatureSets": Empty folder where feature set files will be stored. This folder is referred to as RESDIR in the code.
   - "FeatureCreation": Contains Python and R scripts that create feature set files and put them under RESDIR
   - "Models": Contains R scripts that create the 8 models that went in to our best ensemble submission.
   - "FeatureCreation_Toplevel.r": The R script that creates all features
   - "EnsembleCombination.r": The R script that builds, runs, and merges 8 models into an ensemble.

2. Create Feature sets

   (a) In FeatureCreation_toplevel.r, change the pointers MAIN_DIR, and DATADIR to point to appropriate directories.

(b) In R, go to MAIN_DR and Run FeatureCreation_toplevel.R. It will create following 22 files in the RESDIR (listed in alphabetical order)

```
[1] "edit_times.csv"
 [2] "edit_times_unrounded.csv"
 [3] "editor_list.txt"
 [4] "editor_monthly_edits.csv"
 [5] "editor_monthly_edits_unique_days.csv"
 [6] "editor_monthly_edits_unique_sessions_unrounded.csv"
 [7] "editor_monthly_edits_unrounded.csv"
 [8] "editor_regdate.csv"
 [9] "eid_fsdt_regdate_table.csv"
[10] "Features113_XinP1P2_YinP3.csv"
[11] "Features113_XinP1P2P3.csv"
[12] "Featureskd_wrevs_XinP1P2_YinP3.csv"
[13] "Featureskd_wrevs_XinP1P2P3.csv"
[14] "Featureskd_XinP1P2_YinP3.csv"
[15] "Featureskd_XinP1P2P3.csv"
[16] "leaderboard_features.csv"
[17] "OrigSet_xinp1p2_yinp3.csv"
[18] "OrigSet_xinp1p2p3.csv"
[19] "OrigSubset1_Lead.csv"
[20] "OrigSubset1_Train.csv"
[21] "reverts_related_features_training_and_LB.csv"
[22] "training_features.csv"
```

On a typical machine with 2 GB RAM, this process takes 10 to 12 hours.

3. Run “EnsembleCombination.R” to generate the final result file “Pick3.csv”. The file “EnsembleCombination.R” runs all the 8 constituent models and generates the final result.

For any difficulty running the files please contact
Roopesh : roopesh.ranjan@gmail.com
or Kalpit : itskalpit@gmail.com