# Data Drop: Extra Features

This folder provides several features that may be helpful -- feel free to utilize these! Of course, you may already have variants of these features, or have created other features that captured the corresponding signals. Good luck!

* **ClusterFeatures.txt, ClusterFeatures1.txt**These features were produced by a clustering algorithm in [DebugAdvisor](http://research.microsoft.com/apps/pubs/default.aspx?id=101301) system, which groups together similar bugs.   Each line contains bug id and corresponding cluster id for bugs that were grouped together.   Some bugs remained in their own (singleton) clusters, and are not included.

Hint for utilizing the features: the cluster identity can be added as a categorical feature, with the not-present “singleton” clusters getting a singular “rest” feature.

* **LDAtopics\***

These features were produced by running [Latent Dirichlet Allocation](http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation) topic modeling on the bugs, and associating each bug with a vector of obtained topics. Each of the 40 columns corresponds to the membership probability for the 40 topics constructed by LDA.

* **ZTrace100\***

These features were produced by taking the last 100 lines of the ZTrace associated with the bug, and doing basic text normalization that removes punctuation and casing. This results in a “bag of words” (or unique words for \*unique\*.tsv).

Hint for utilizing the features: in TLC, these can be readily plugged into a Text handler or TextHash handler.