Automatic Log Analysis using Machine Learning
Applies machine learning techniques to do automated log analysis. Compares several variants of clustering, artificial neural network algorithms and data preprocessing.
http://uu.diva-portal.org/smash/get/diva2:667650/FULLTEXT01.pdf

Overview of the paper:
Published Nov 2013 – some time ago, maybe it's worth to look for something newer

Working on unstructured text logs, mixed and single configuration (logs from different sources) - as in DPCS

Text preprocessing tips:
1) Replace timestamps. Use a special symbol to replace the whole timestamp before each message.
2) Replace digits. Use a special symbol to replace any digit in the log file.
3) Lower cases. Change all upper case letters into lower case.
4) Remove special characters. Remove all special characters, including punctuations, and only keep letters and digits.

Features:
Manually created, shortage on expert knowledge
Char bigram, word bigram, word count, timestamp stats (different metrics on differences between subsequent timestamps in the log)
TF (term frequency - normalised wc) + IDF (importance of the word - how frequently is it used between logs)

Clustering:
Two classes (anomaly detection), not very useful in our case
DBSCOD - density based spatial clustering of outliers detection
core point - If the number of points in one point p's neighbourhood is greater than the threshold MinPts, p is a core point.
border point - The border point is not a core point, but it is located in one or multiple core points' neighbourhoods.
outlier - (noise point) The other points except the core points and border points are all utliers.
They were interested in outliers, as they were probably anomalies.

Self-organising feature maps - simple explanation from wiki
https://en.wikipedia.org/wiki/Self-organizing_map#Learning_algorithm

Results:
K-means and other simple algorithms doing terrible, SOFM is good
High score on mixed configuration (different types of logs using one classifier)
Better to do dimensionality reduction for whole dataset (features * attributes) rather than separately for every feature

Important note about features:
"Secondly, among those feature candidates, the results show that simple features such as character bigram and timestamp statistics are effective enough to distinguish abnormal and normal logs. The advanced feature TF-IDF is more effective in certain test case, but gets fair results in some other test cases."