

## Log Parsing

Example log lines from the Linux dataset containing month, date, time, log level, logging component, and the log message (bold):

```
Oct 23 12:40:02 combo cups-lpd[22514]: Unable to get command line from client!
Oct 25 10:08:47 combo kernel: Inode-cache cache hash table entries: 16384 (order: 4, 65536 bytes)
Nov 22 14:31:58 combo kernel: httpd: page allocation failure. order:0, mode:0x1d2
Nov 22 14:32:24 combo kernel: sendmail: page allocation failure. order:0, mode:0x1d2
Dec 6 12:22:57 combo kernel: PID hash table entries: 512 (order: 9, 4096 bytes)
```



Templates resulting from log parsing. Underlined parts indicate the desired MWEs:

```
Unable to get command line from client!
<*> hash table entries: <*> (order: <*>, <*> bytes)
<*>: page allocation failure. order:<*>, mode:<*>
```

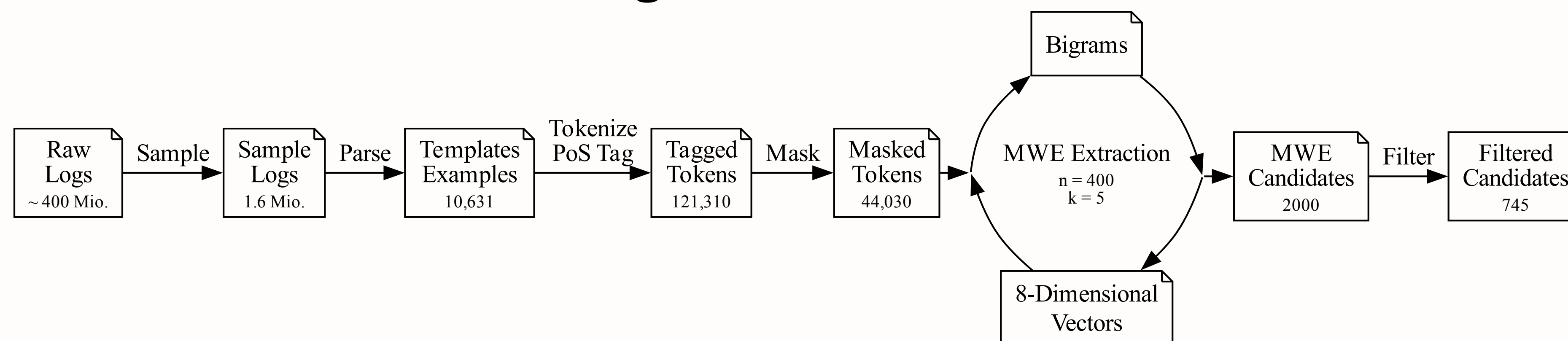
## The Dataset

We use the 16 datasets from Loghub [1]. They represent a wide range of software systems, such as

- distributed systems,
- supercomputers,
- operating systems,
- mobile systems,
- server applications, and
- standalone software.

In total, these datasets contain over 400 million logs.

## From Logs to MWE Candidates



Our automated data processing pipeline applying Gries' MWE extraction method [2].

## Majority Vote

Proportions of successful and unsuccessful MWE candidates with majority vote by three annotators.

MWE Candidates	Count	Rel.
Unsuccessful	253	34 %
Successful	492	66 %
Total	745	100 %

## PoS-Filter

Only keep nouns, proper nouns, adjectives and adverbs (tagged with PosLog [3]). A technical term built from adjective and noun would be *public key*, *cryptographic api* or *red hat*, for example.

## Post-Filter

1. Punctuation characters (e.g., *ccfile::copyfile, krbtgt/#24#@#24#*),
2. Tokens with more than 15 characters (e.g., *ksserverupdaterequestdelegate*), and
3. Last token does not have the PoS tag noun or proper noun (e.g., *too many, so far*).

## Iteratively Growing MWEs

Examples from the listing above:

- *hash table* (iteration 22),
- *hash table entries* (iteration 47),
- *page allocation* (iteration 64),
- *page allocation failure* (iteration 77),
- *command line* (iteration 159).

## Results

### Classification

Classification results on successful MWE candidates with majority vote. Parity arises when two annotators voted for successful, but different classes.

Class	Count	Rel.
Technical Terms	424	86 %
Proper Nouns	30	6 %
Majority Vote	454	92 %
Parity	38	8 %
Total	492	100 %

Proper noun examples:

- *red hat linux* (Company, Thunderbird),
- *dave jones* (Person, Linux),
- *internet systems consortium* (Group, Thunderbird).

Note: Single-token proper nouns such as *Linux* or *Google* are no MWE.

### Occurrence

Successful MWE occurrences per dataset.

Dataset	Count	Rel.
Thunderbird	490	26.36 %
Android	361	19.42 %
BGL	342	18.40 %
Mac	305	16.41 %
Linux	168	9.04 %
Hadoop	70	3.77 %
Windows	52	2.80 %
Spark	17	0.91 %
OpenSSH	16	0.86 %
HPC	13	0.70 %
Zookeeper	10	0.54 %
Proxifier	5	0.27 %
OpenStack	4	0.22 %
HDFS	3	0.16 %
HealthApp	2	0.11 %
Apache	1	0.05 %
Total	1,859	100 %

Combining MWE tokens saves 1,859 of the 121,310 total tokens. This results in a savings rate of 1.5 %.

### Most Common

Top 5 most common in all datasets:

- *image pages* (41x),
- *update check* (18x),
- *button report* (18x),
- *authentication failure* (15x),
- *hash table* (14x).

### Most Widely Spread

Top 5 most spread across the datasets:

- *state change* (in 4 datasets),
- *file descriptor* (in 4 datasets),
- *authentication failure* (in 3 datasets),
- *configuration file* (in 3 datasets),
- *host controller* (in 3 datasets).

### Longest Terms

Distribution of successful MWE candidates by their length in tokens.

Tokens	Count	Rel.
2	362	79.7 %
3	71	15.6 %
4	13	2.9 %
5	5	1.1 %
6	1	0.2 %
Total	454	100.0 %

- *google software update installer* (Mac),
- *pci hot plug pci core* (Linux),
- *usb universal host controller interface driver* (Thunderbird).

## References

- [1] Jieming Zhu et al. "Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics". In: *IEEE International Symposium on Software Reliability Engineering (ISSRE)*. 2023.
- [2] Stefan Th. Gries. "Multi-Word Units (and Tokenization More Generally): A Multi-Dimensional and Largely Information-Theoretic Approach". In: *Lexis* 19 (Mar. 2022).
- [3] Kilian Dangendorf et al. "PosLog: Creating a Part of Speech Tagger for Log Messages". In: *2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. 2025, pp. 1444-1449. doi: 10.1109/IDAACS68557.2025.11322035.

## Let's Connect



LinkedIn