System Log Glossary Search (SLGS)

CSC482 - Daniel Gonzalez

Background

Engineers often test code on ephemeral environments (CI) to ensure code compliance, though during these tests the system may vary in architecture and resources which may cause unexpected behavior, occurrences like this may occur in multiple pull-requests and make hinder the developer experience

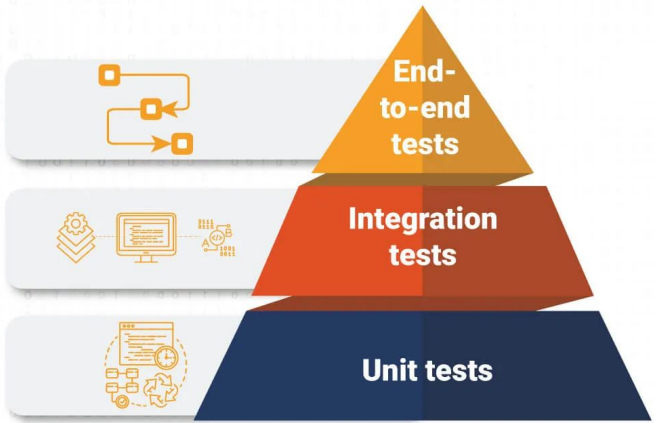


Figure 1. Testing Hierarchy

Original DataSet

Dataset

- https://zenodo.org/records/4008017
- Raw Data for Experiments of Automated Evolution of Logging Statement Levels Using Git Histories and Degree of Interest

sequence	subject raw	repo URL	decay factor	input logging statements	candidat
1579726998335	IRCT-API	https://github.com/hms-dbmi/IRCT.git	256	9	
1579726998335	IRCT-CL	https://github.com/hms-dbmi/IRCT.git	256	0	
1579726998335	IRCT-EXT	https://github.com/hms-dbmi/IRCT.git	256	0	
1579726998335	IRCT-RI	https://github.com/hms-dbmi/IRCT.git	256	5	
1579726998335	SciDB-Java	https://github.com/hms-dbmi/IRCT.git	256	0	
1579726998335	i2b2-Java-API	https://github.com/hms-dbmi/IRCT.git	256	0	
1579731072210	opengrok	https://github.com/oracle/opengrok.git	256	495	
1579731072210	opengrok-web	https://github.com/oracle/opengrok.git	256	44	
1579731072210	plugins	https://github.com/oracle/opengrok.git	256	43	
1579731072210	suggester	https://github.com/oracle/opengrok.git	256	44	
1579776483488	bc-java	https://github.com/bcgit/bc-java.git	256	56	
1579876000997	CoreNLP	https://github.com/stanfordnlp/CoreNLP.git	256	409	
1579989136879	error_prone_annotation	https://github.com/google/error-prone	256	0	
1579989136879	error_prone_annotations	https://github.com/google/error-prone	256	0	
1579989136879	error_prone_check_api	https://github.com/google/error-prone	256	6	
1579989136879	error_prone_core	https://github.com/google/error-prone	256	7	
1579989136879	error_prone_docgen	https://github.com/google/error-prone	256	0	
1579989136879	error_prone_docgen_processor	https://github.com/google/error-prone	256	0	

Figure 2. Sample Logs

Fetching from 50 GitHub Repo

GitHub Scrape

- Fetched 50 GH's
- Regex relevant logs

global_logs

	TYPE	LOG
f1984c45ff439.zip	WARN	2023-10-04T19:34:14.8368590Z WARNING | encryption is off
f1984c45ff439.zip	WARN	2023-10-04T19:34:14.8457240Z WARNING | Running on a system with less than 6 logical cores. Setting number of virtual cores to 1
f1984c45ff439.zip	ERROR	2023-10-04T19:34:15.3697940Z ERROR | Unable to connect to adb daemon on port: 5037
f1984c45ff439.zip	WARN	2023-10-04T19:34:16.3276750Z WARNING | cannot add library /Users/runner/Library/Android/sdk/emulator/qemu/darwin-x86_64/lib64/vulkan/libvu
f1984c45ff439.zip	WARN	2023-10-04T19:34:16.4425770Z WARNING | /etc/localtime does not point to zoneinfo-compatible timezone name
f1984c45ff439.zip	WARN	2023-10-04T19:34:16.4451790Z WARNING | *** No gRPC protection active, consider launching with the -grpc-use-jwt flag.***
f1984c45ff439.zip	WARN	2023-10-04T19:34:16.4843590Z WARNING | Failed to process .ini file /Users/runner/.android/emu-update-last-check.ini for reading.

```
top_java_repos = [
    "google/guava", # 25 most starred repos
    "CyC2018/CS-Notes",
    "Snailclimb/JavaGuide",
    "iluwatar/java-design-patterns",
    "doocs/advanced-java",
    "macrozheng/mall",
    "spring-projects/spring-boot",
    "elastic/elasticsearch",
    "kdn251/interviews",
    "TheAlgorithms/Java",
    "azl397985856/leetcode",
    "google/guava",
    "ReactiveX/RxJava",
    "square/okhttp",
    "youngyangyang04/leetcode-master",
    "square/retrofit",
    "apache/dubbo",
    "apache/spark",
    "skylot/jadx",
    "PhilJay/MPAndroidChart",
    "jeecgboot/jeecg-boot",
    "autonomousapps/dependency-analysis-gradle-plugin", # common gradle plugins
    "modrinth/minotaur",
    "klawson88/liquiprime",
    "klawson88/liquigen",
```

Figure 3. Repository List

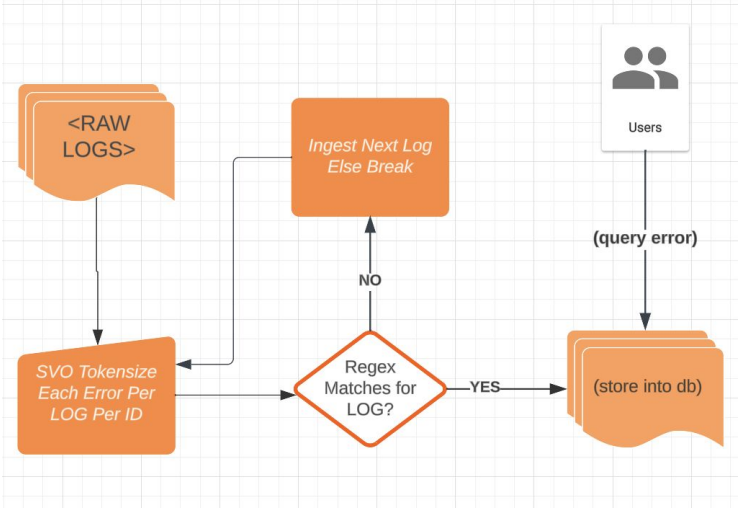


Figure 4. DFD & Logs

SVO Extraction/Testing

Implementation

- Parse SVO's using SpaCy
- Provide log files via CLI and returned best matched

	SVO
ID	SVOs
logs_square-retrofit-6410640346-27618cf62ffbfa037f96389cd52f1984c45ff439.zip	['encryption', 'system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'c
logs_mockito-mockito-6547057197-d74808b2510125c1b44460149d17c7eabb7a3db5.zip	['system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'cores', 'daemo
logs_square-retrofit-6209378482-c5adac55c531390155812cb8b9b3ab710c936c1a.zip	['encryption', 'system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'c
logs_mockito-mockito-6776473606-32bf2d3e85e5035db95669b6af4d27bab0d41085.zip	['system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'cores', 'daemo
logs_square-retrofit-6476395622-f7c3b3e299ee2bf07fa361d7b19a19e909b373f4.zip	['encryption', 'system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'c
logs_square-okhttp-6925087984-7db40afae85f446c5a9515ef018e45a0e62ecdc4.zip	['system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'cores', 'daemo
logs_square-okhttp-6452058572-3e9ea685b2946f395e21401f123405aeb366d38a.zip	['system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'cores', 'daemo
logs_square-okhttp-6527663963-ced8465c1b6791cdc9149a1e37c17b2587338ab4.zip	['encryption', 'system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'c
logs_mockito-mockito-7050940489-97200b14a2d19a776f777532b834cd0a27f8f691.zip	['system', 'cores', 'number', 'cores', 'system', 'cores', 'number', 'cores', 'daemo

Figure 5. Parsed SVO Sample

```
→  CSC482-Log-NLP git:(main) python3 searchSVO.py sample_log_1.txt
From 1812 SVO matches, here is the link to a commit where the builds matched the errors provided:
https://github.com/square/okhttp/commit/4984568367caaf359b82c452bd28b5e192824d1c
→  CSC482-Log-NLP git:(main) python3 searchSVO.py sample_log_2.txt
From 209 SVO matches, here is the link to a commit where the builds matched the errors provided:
https://github.com/square/okhttp/commit/4984568367caaf359b82c452bd28b5e192824d1c
```

Figure 6. CLI Output

Java OSS Travis-CI Build Failure Dataset

Dataset

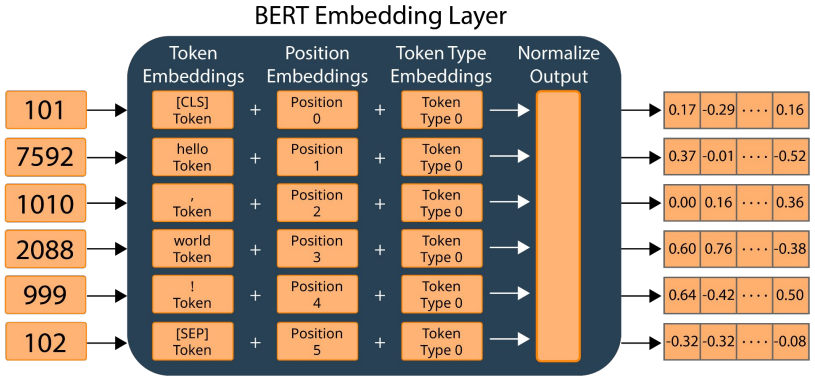
- https://zenodo.org/records/1745638
- Resulted in 85597 txt files (each a CI build with errors)

- **git-repositories.tar.gz**
  contains the Git repositories analyzed. Each commit that triggered a Travis-CI build (beginning from the time the continuous mining process was started) is tagged with the Travis-CI build ID
- **logcat-categories-and-patterns.tar.gz**
  contains the error categories of the analyzed log files that were coded from the exploratory analysis
- **projects.csv**
  maps the Travis-CI project ID to the GitHub project slug
- **travis-ci-logs.tar**
  contains for each project the analyzed log files that were scraped from Travis-CI. Each txt.gz file in the project folders contains the log of one Travis *job*. A build may have multiple jobs.

Figure 7. Travis-CI Data Provided

Bert Embeddings

- Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.



ID	embeddings
data/1129565/10527996	[0.2890910208225
data/1129565/10544933	[0.2890910208225
data/1129565/10548521	[0.2890910208225
data/1129565/10549679	[0.2890910208225
data/1129565/10576070	[0.2890910208225
data/1129565/10576578	[0.2890910208225
data/1129565/10582858	[0.2890910208225
data/1129565/10583282	[0.2890910208225
data/1129565/10583356	[0.2890910208225
data/1129565/10583928	[0.0501867271959781
data/1129565/10591670	[-0.218967005610466
data/1129565/10594935	[-0.218967005610466
data/1129565/10595860	[-0.218967005610466
data/1129565/10911150	[-0.218967005610466
data/1129565/10923401	[-0.218967005610466
data/1129565/11046616	[0.2890910208225

Figure 8. Bert Embedding System & Sample Output

Conclusion

Results Matrix

Cosine Similarity	LOG COUNT (N = 25%)
1.0000	100
0.9998	1000
0.9837	5000
0.9711	50000

Figure 9. Results Matrix

Future Improvements:

- Preprocessing, errors can arise in various structures, Gradle, SL4J, and LOG4J each in their own format. Preprocessing for each, would reduce the data size and also be a helpful abstraction for the model to ingest from.
- Fine-tuning the BERT model on log-specific data or using domain-specific embeddings if available. LogBERT For example: https://arxiv.org/abs/2103.04475
- Implement optimizations for better performance, especially when dealing with a large number of logs. 4.4 Gigabytes of CSV is difficult to process independently and can be better distributed.