Build log analysis for error detection and deduplication

Irving Muller Rodrigues
Rodrigo Alves Randel
Mohammad Nassiri

Polytechnique Montréal
Laboratoire DORSAL

Progress Report Meeting, January 21, 2022

Agenda

1. Background – build automation system
   - Build log errors
2. Motivation and Problem Statement
3. Extendable Framework
4. Identification and Correlation of Python Errors
   - Pre-processing Python tracebacks
   - A new similarity measure
   - Deduplication using clustering
   - Performance improvement
5. Conclusion and Future work

Background – build automation system

- Build automation system is a part of DevOps workflow and involves scripting or automating the process of compiling, linking and testing computer source code into binary code.
- Some tools for build automation: make, Ninja, CMake, Meson, ...
- Our use case is based on make and the build automation is a stage in Jenkins pipeline system.
  - Hierarchical Python scripts and Makefiles
- Build outcome: Fail or Success

Example of Errors in a Failed Build Log

```
In file included from /path3/hfile3.h:108,
                 from /path2/target/hfile2.h:38,
                 from /path1/sources/hfile1.h:28,
                 from main.c:8:
/path3/hfile4.h:200: stdarg.h: No such file or directory
/path3/hfile4.h:300: error: syntax error before "list_1"
```

Compile error

```
make[2]: *** [mytarget.o] Error 1

make: *** [all] Error 1

make: *** [main_data.c] Error 255

make: ccArch: Command not found
```

Make error

```
XQuery error (255) processing /localdisk/XXX/data_model/abc2def.xqy
XQuery error (255) processing /localdisk/XXX/data_model/abc2mno.xqy
```

XQuery error

```
Traceback (most recent call last):
   File "/localdisk/tools/package/deploy", line 60, in <module>
     pkg.deploy()
   File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
     process(artifact_d, version)
   File "/localdisk/tools/package/src/__init__.py", line 300, in process
     rt.deploy(src, artifact)
   File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
     with open(src) as f:
IOError: [Errno 2] No such file or directory: u'/localdisk/build/product.po'
```

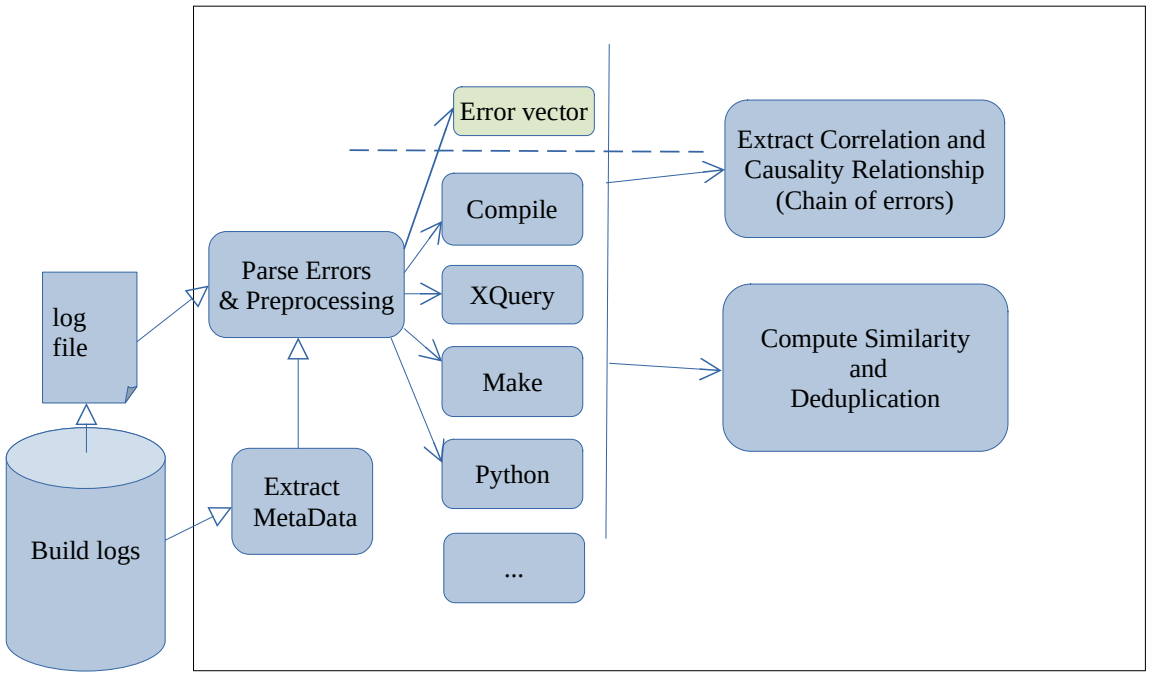
Python error

- One or more patterns for each error type in the build log

Motivation and Problem Statement

- Multiple daily releases each including hundreds of different build targets
- Identify and correlate build errors within one build and across multiple builds
- Same issue may have different manifestations in different builds
- Various errors in a single build may all be linked back to a single root cause
- Special case is error deduplication

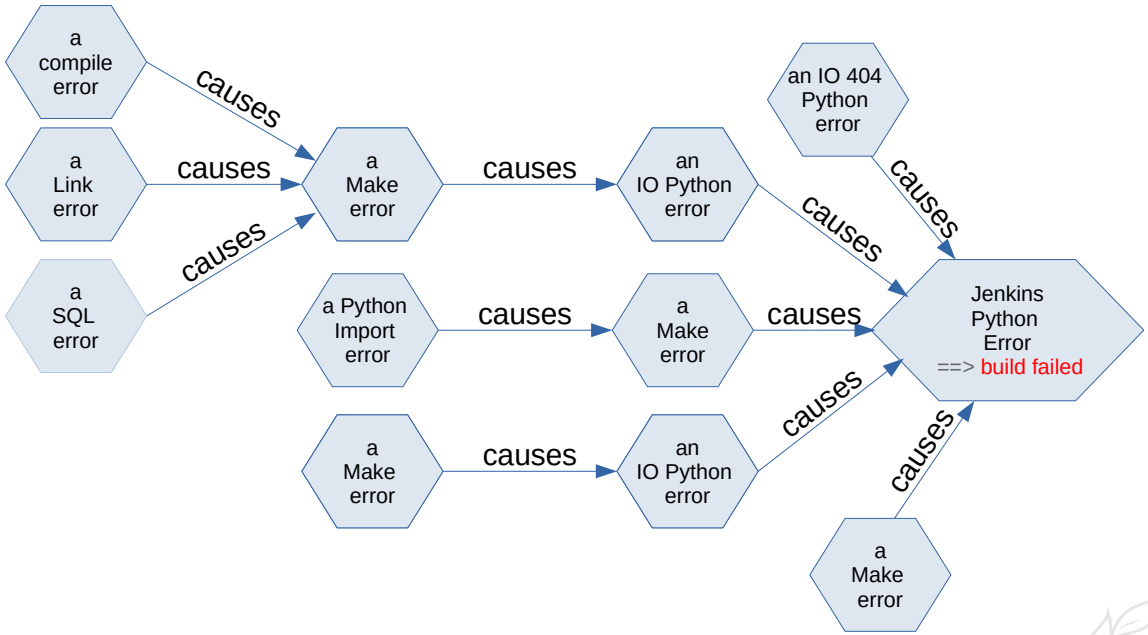
Extendable Framework



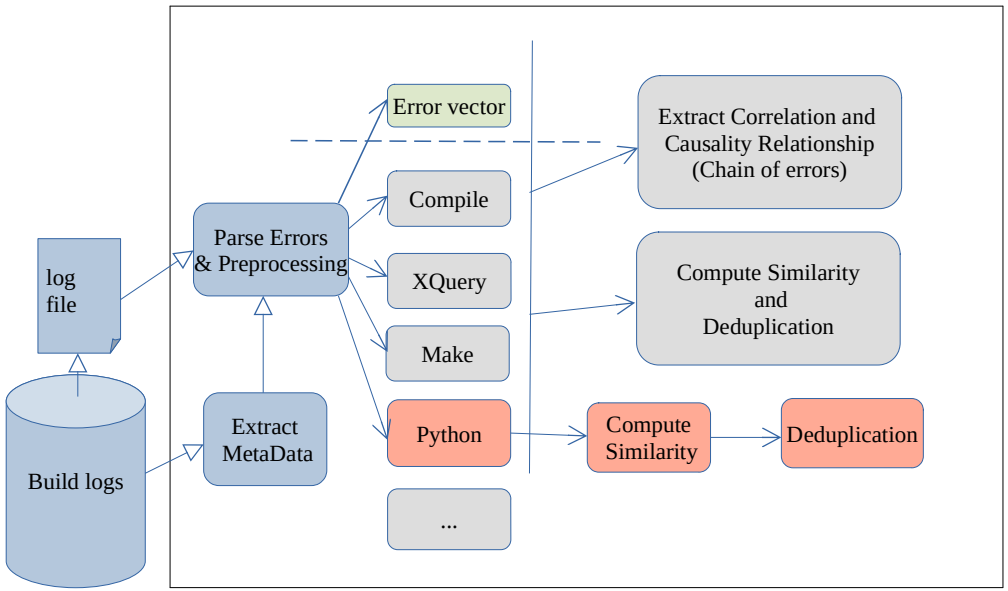
- Extract metadata from build logs
- Parse and vectorize build errors
- Analyze error vectors using model
  - Similarity and deduplication across builds
  - Causality in a single build

Correlate Errors Within Single Build

- Multiple build errors tend to be linked back to single root cause
- De-clutter build errors by finding causality chain
- Reduce debugging time



Framework



- Rodrigo to present the Python-related part
- Vithor: our new PhD student to work on
  - Using machine learning to find causality between errors in log reports.
  - Presentation at 9:55-10:10

Build log mining - Python Errors

Error A

```
Traceback (most recent call last):
        File "/localdisk/tools/package/deploy", line 60, in <module>
            pkg.deploy()
        File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
            process(artifact_d, version)
        File "/localdisk/tools/package/src/__init__.py", line 300, in process
            rt.deploy(src, artifact)
        File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
            with open(src) as f:
        IOError: [Errno 2] No such file or directory: u'/localdisk/build/target1/AA/platformX/productA.po'
```

Error B

```
Traceback (most recent call last):
        File "/localdisk/tools/package/deploy", line 60, in <module>
            pkg.deploy()
        File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
            process(artifact_d, version)
        File "/localdisk/tools/package/src/__init__.py", line 300, in process
            rt.deploy(src, artifact)
        File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
            with open(src) as f:
        IOError: [Errno 2] No such file or directory: u'/localdisk/build/target2/BB/platformY/productB.po'
```

Build log mining - Python Errors

Error A

```
Traceback (most recent call last):
    File "/localdisk/tools/package/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target1/AA/platformX/productA.po'
```

Error B

```
Traceback (most recent call last):
    File "/localdisk/tools/package/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target2/BB/platformY/productB.po'
```

S_{frames}

Build log mining - Python Errors

Error A

```
Traceback (most recent call last):
        File "/localdisk/tools/package/deploy", line 60, in <module>
            pkg.deploy()
        File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
            process(artifact_d, version)
        File "/localdisk/tools/package/src/__init__.py", line 300, in process
            rt.deploy(src, artifact)
        File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
            with open(src) as f:
        IOError: [Errno 2] No such file or directory: u'/localdisk/build/target1/AA/platformX/productA.po'
```

Error B

```
Traceback (most recent call last):
        File "/localdisk/tools/package/deploy", line 60, in <module>
            pkg.deploy()
        File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
            process(artifact_d, version)
        File "/localdisk/tools/package/src/__init__.py", line 300, in process
            rt.deploy(src, artifact)
        File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
            with open(src) as f:
        IOError: [Errno 2] No such file or directory: u'/localdisk/build/target2/BB/platformY/productB.po'
```

S_{frames}

S_{msg}

Build log mining - Python Errors

Error A

```
Traceback (most recent call last):
    File "/localdisk/tools/package/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target1/AA/platformX/productA.po'
```

Error B

```
Traceback (most recent call last):
    File "/localdisk/tools/package/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target2/BB/platformY/productB.po'
```

S_{frames}

S_{msg}

S_{url}

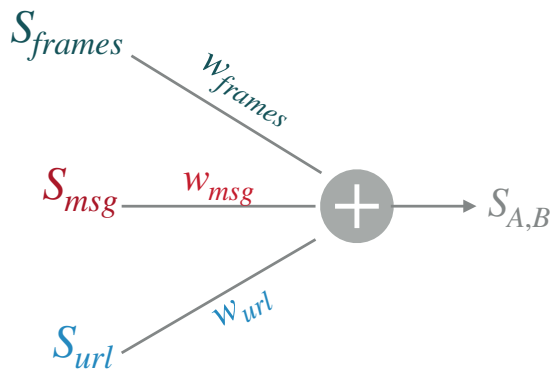
Build log mining - Python Errors

Error A

```
Traceback (most recent call last):
    File "/localdisk/tools/package/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target1/AA/platformX/productA.po'
```

Error B

```
Traceback (most recent call last):
    File "/localdisk/tools/package/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target2/BB/platformY/productB.po'
```



Python Errors

Preprocessing the traceback

Error A

```
Traceback (most recent call last):
    File "/localdisk/tools/package/target1/AA/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target1/AA-202/platformX/productA.po'
```

Error B

```
Traceback (most recent call last):
    File "/localdisk/tools/package/target2/BB/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/target2/BB-105/platformY/productB.po'
```

Python Errors

Preprocessing the traceback

Error A

```
Traceback (most recent call last):
    File "/localdisk/tools/package/<target>/<release>/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/<target>/<release>-<version>/<platform>/productA.po'
```

Error B

```
Traceback (most recent call last):
    File "/localdisk/tools/package/<target>/<release>/deploy", line 60, in <module>
        pkg.deploy()
    File "/localdisk/tools/package/src/__init__.py", line 200, in deploy
        process(artifact_d, version)
    File "/localdisk/tools/package/src/__init__.py", line 300, in process
        rt.deploy(src, artifact)
    File "/localdisk/tools/package/src/artifactory/__init__.py", line 100, in deploy
        with open(src) as f:
    IOError: [Errno 2] No such file or directory: u'/localdisk/build/<target>/<release>-<version>/<platform>/productB.po'
```

Python Errors

Similarity of error messages

Event A

```
IOError: [Errno 404] File not found
```

TF-IDF

Word 1	Word 2	...	...	...	...	...	Word n
0	0.145	...	0.278	...	0.79	...	0

Event B

```
UnboundLocalError: local variable 'r' referenced before assignment
```

TF-IDF

Word 1	Word 2	...	...	...	...	...	Word n
0	0	0.89	...	0.36	...	0.58	0

Cosine Similarity

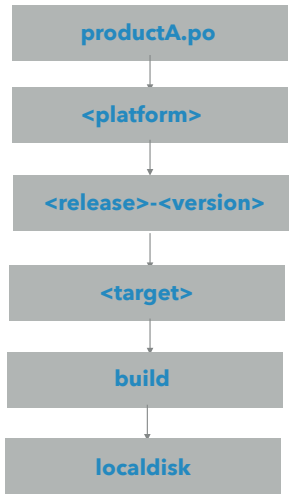
$$S_{msg}(E_A, E_B) = 0.0$$

Python Errors

Similarity of URLs

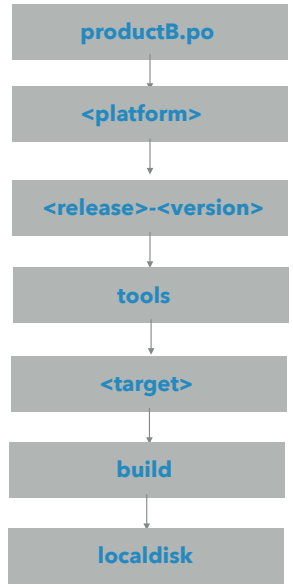
Event A

URL: /localdisk/build/<**target**>/<**release**>-<**version**>/<**platform**>/productA.po



Event B

URL: /localdisk/build/<**target**>/tools/<**release**>-<**version**>/<**platform**>/productB.po



Weighted global alignment

$$S_{url}(E_A, E_B) = 0.775$$

Python Errors

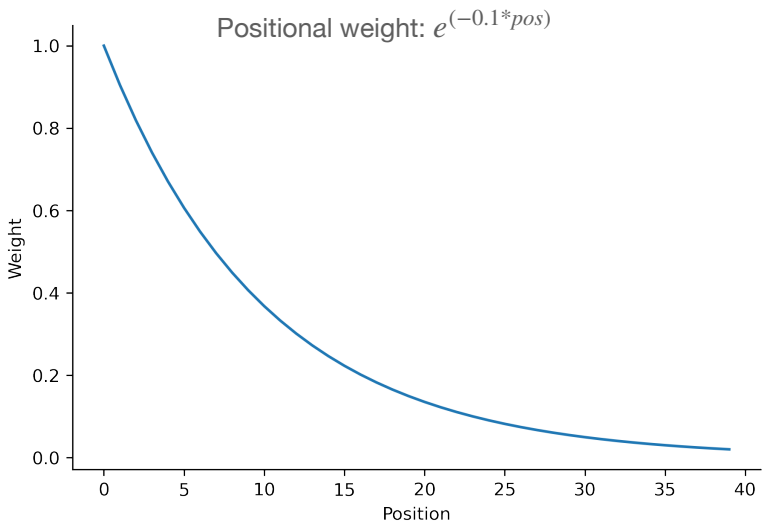
Similarity of URLs

URL A	productA.po	<platform>	<release>-<version>		<target>	build	localdisk
URL B	productB.po	<platform>	<release>-<version>	tools	<target>	build	localdisk

Python Errors

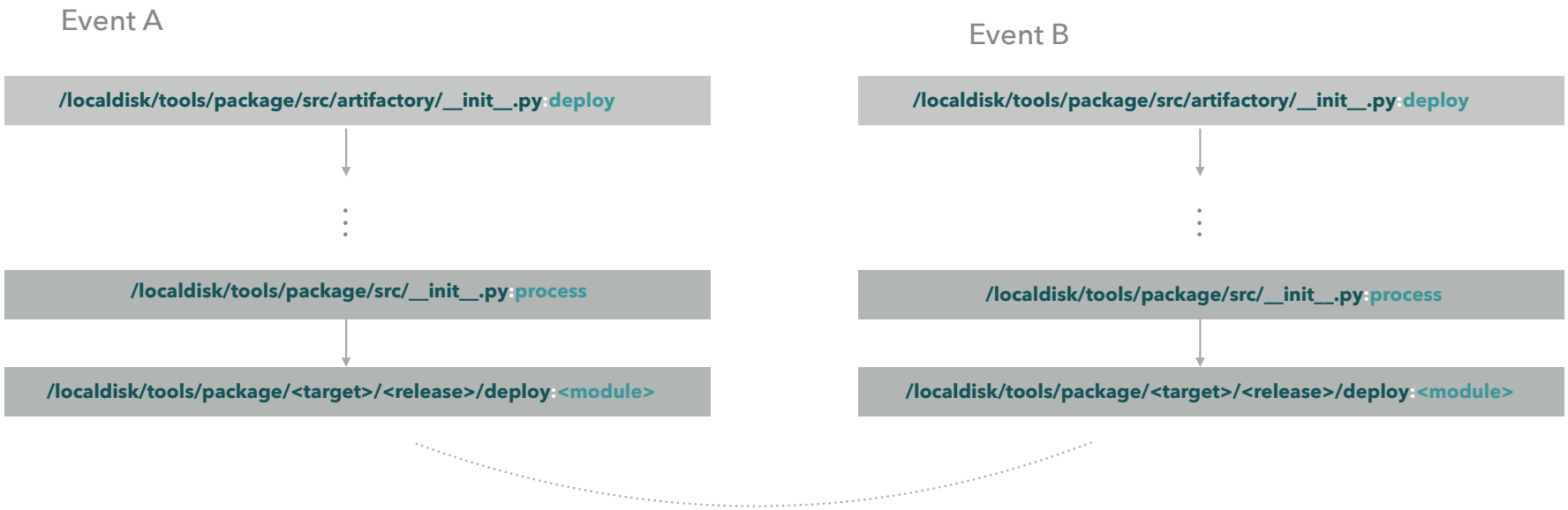
Similarity of URLs

URL A	productA.po	<platform>	<release>-<version>		<target>	build	localdisk
URL B	productB.po	<platform>	<release>-<version>	tools	<target>	build	localdisk



Python Errors

Similarity of Frames



$$S_{frames}(E_A, E_B) = 1.0$$

Results for Python Errors

About the data

- 509 build logs
- 174 Python Errors

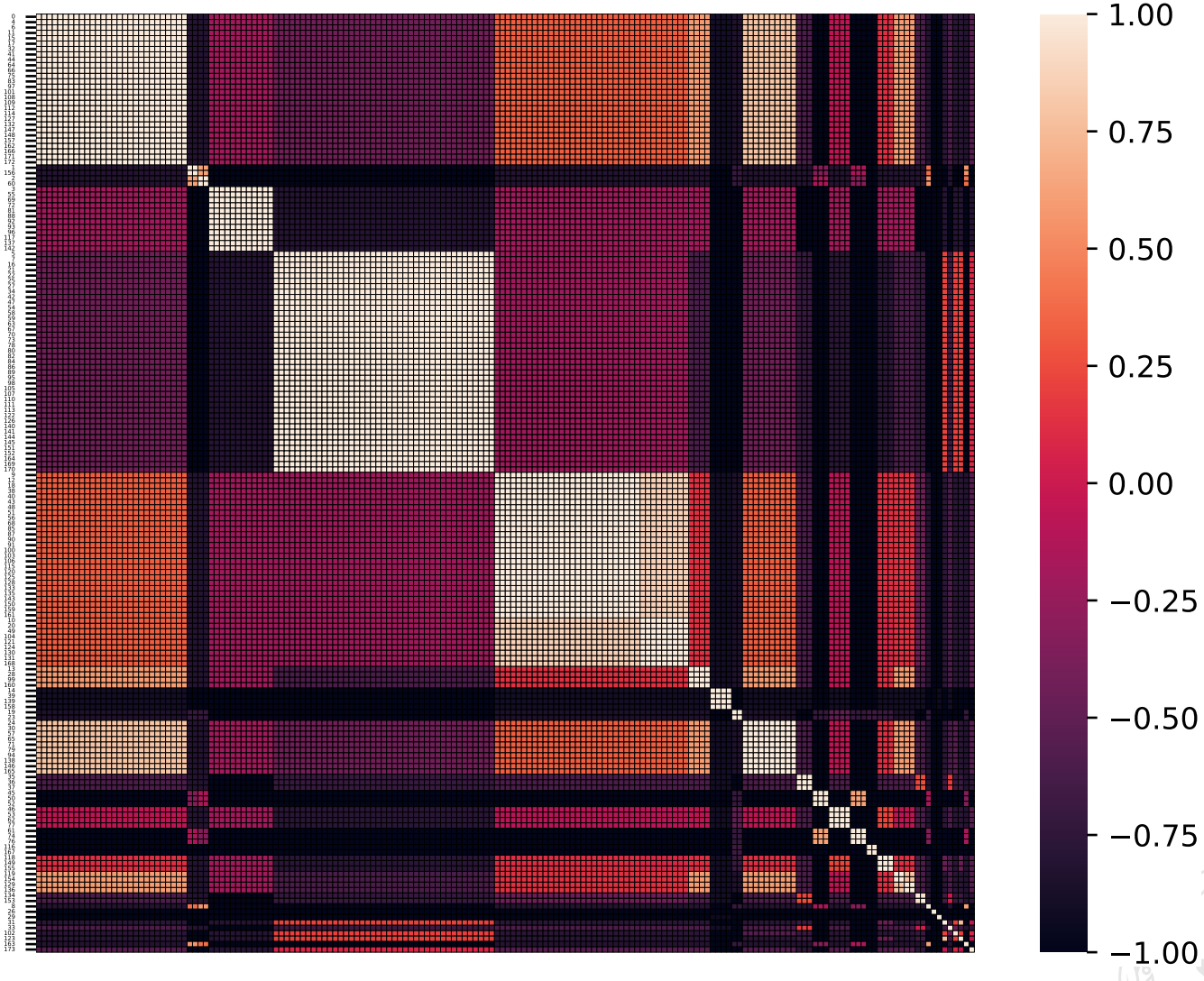
Goals

- Achieve good deduplication performance

Tuning w_{frames}, w_{url}, w_{msg}

- Clustering analysis with external validation (company)

Clustering for deduplication - DBSCAN



Performance improvement

Time became an issue

- The new python similarity was significantly augmenting time to process build logs
- Provoking delays on the workflow

- Re-implemented the Python Traceback Similarity using Cython

Python

```python
for (q_pos, q_func) in enumerate(query):
    # Set first column of the row
    diagonal = row[0]
    row[0] -= weights[q_pos]

    for (c_pos, c_func) in enumerate(candidate):
        col_idx = c_pos + 1

        # Align gap to query position
        above = row[col_idx] - weights[q_pos]

        # Align gap to candidate position
        left = row[col_idx - 1] - weights[c_pos]

        if q_func == c_func:
            # Matching
            diagonal += weights[c_pos] + weights[q_pos]
        else:
            # Mismatch
            diagonal -= 2 * (weights[c_pos] + weights[q_pos])

        diagonal_tmp = row[col_idx]

        # Replace max with IF ELSE considerably improves the performance =0
        row[col_idx] = max(left, above, diagonal)

        diagonal = diagonal_tmp
```

Cython

```cython
cdef char * q_func
cdef char * c_func
cdef float diagonal, above, left, diagonal_tmp
cdef int col_idx, q_pos, c_pos
for q_pos in range(q_len):
    q_func = query[q_pos]
    # Set first column of the row
    diagonal = self.row[0]
    self.row[0] = self.row[0] - self.weights[q_pos]

    for c_pos in range(cand_len):
        c_func = candidate[c_pos]
        col_idx = c_pos + 1

        # Align gap to query position
        above = self.row[col_idx] - self.weights[q_pos]

        # Align gap to candidate position
        left = self.row[col_idx - 1] - self.weights[c_pos]
        if strcmp(q_func,c_func) == 0:
            # Matching
            diagonal += self.weights[c_pos] + self.weights[q_pos]
        else:
            # Mismatch
            diagonal -= 2 * (self.weights[c_pos] + self.weights[q_pos])

        diagonal_tmp = self.row[col_idx]

        # Replace max with IF ELSE considerably improves the performance =0
        self.row[col_idx] = max(left, above, diagonal)

        diagonal = diagonal_tmp
```

Figure: Python × Cython illustration

- Reduced the processing time from $\sim 225s$ to $\sim 2.3s$ ($\sim 100x$ speed up)

Future works

- Develop/extend the parser to extract any error types
- Identify the chain of errors for each build log
  - The tail element in the chain could be the root cause error
  - Investigate association rules to extract similar chain of errors

Any questions/comments?
Thank You