Improving DevOps and QA efficiency using machine learning and NLP methods

Omer Sagi

May 2018
Outline

- Investigating failed tests
- Document similarity
- The developed search system
  - Unsupervised similarity
  - Supervised similarity learning
- Evaluation
The business domain – Investigating failed tests
Investigating failed tests - overview

New test design → Failed test → Test investigation

Log file

Tracking

JIRA, Bugzilla, RALLY, slack
Investigating failed tests – duplicated issues

Ticketing system / Issue-26626

**Issue:** Test was failed due to network connectivity issue

<table>
<thead>
<tr>
<th>Issue Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>duplicates</td>
</tr>
<tr>
<td><code>issue-26626</code> Data unavailability due to system <code>xxx</code> shutdown</td>
</tr>
<tr>
<td>relates to</td>
</tr>
<tr>
<td><code>issue-26145</code> DU with auto-recovery</td>
</tr>
<tr>
<td><code>issue-4479</code> Network failure has been detected following system <code>x</code>...</td>
</tr>
</tbody>
</table>

### Activity

- John Doe added a comment - 13/Mar/16 11:39 AM Richard Roe [X] please take a look...
- Richard Roe added a comment - 14/Mar/16 10:03 AM

Please have a look at the following messages that seem to relate issue-26626:

- `<INFO> 2016-02-19 08:08:00 event AAA`
- `<INFO> 2016-02-19 08:08:01 event BBB`
- `<INFO> 2016-02-19 08:08:02 event AAA`
Current state problem

Investigators overlook duplicated issues while mistakenly mark other duplications

Looks like a network outage

Log filesystem

Network outage

GUI issue

Data unavailability

Data loss
Document similarity
Document search system - overview

Query → Text processing → Processed terms → Indexed knowledge base

Ranked Document set

Retrieved document set → Similarity model

Retrieved document set → Relevant document set

Ranking
Documents as vectors

The intuition: Since \([D]\) is a vector we can find other nearby vectors (documents)

Distance between vectors \(d_1\) and \(d_2\) can be *captured* by the cosine of the angle \(\theta\) between them

\[
\text{Cosine similarity } (d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \cdot |d_2|} = 0.66
\]

\[
\text{Cosine similarity } (d_2, d_3) = \frac{d_2 \cdot d_3}{|d_2| \cdot |d_3|} = 0
\]

\begin{array}{cccccc}
\hline
& angeles & los & new & post & times & york \\
\hline
 d1 & 0 & 0 & 1 & 0 & 1 & 1 \\
 d2 & 0 & 0 & 1 & 1 & 0 & 1 \\
 d3 & 1 & 1 & 0 & 0 & 1 & 0 \\
\hline
\end{array}
The developed solution: log-files search system
Solution overview

- Query
- Log processing
- Vector
- Indexed knowledge base
- Relevant log set
- Retrieved issues
- Similarity model
- Ranked issues
Log file vectorization

- Messages are analogous to words, defining the vector dimensionality
- In the following example, each ‘word’ is a concatenation of log type and event id

```plaintext
<info> 00:00:01.344 [event_id 14] system A writes to system B
<crit> 00:00:02.232 [event_id 150] Restarting server xxx
<info> 00:00:02.345 [event_id 23] Start writing file xfgrew21e3456
<error> 00:00:03.232 [event_id 25] failed to write file xfgrew21e3456
<info> 00:00:04.252 [event_id 14] system C writes to system D
```

Vector_dim:
- info_14
- crit_150
- info_23
- error_25

Value:
- 2
- 1
- 1
- 1
Results

Proportion of queries with at least one relevant result

Proportion %

Top 1  Top 3  Top 5  Top 10
Similarity learning

Issue 1

Pair vector

Issue 2

Is duplicate?

Machine-learning model
## Extracting similarity vector

### Issue 1
- **<info> 00:00:01.344 [event_id 1]** Panic 12 - unable to perform update xrt
- **<crit> 00:00:02.232 [event_id 150]** Restarting server xxx
- **<info> 00:00:02.345 [event_id 23]** Start writing file xfgrew21e3456
- **<error> 00:00:03.232 [event_id 25]** failed to write file xfgrew21e3456
- **<info> 00:00:04.252 [event_id 32]** Panic 14 - unable to execute operation mjk

### Issue 2
- **<info> 09:10:43.294 [event_id 1]** Panic 12 - unable to perform update gfds
- **<info> 09:10:45.594 [event_id 65]** User 578903 authentication starts
- **<crit> 09:10:45.742 [event_id 68]** User 578903 authentication stopped
- **<info> 00:00:02.345 [event_id 23]** Start writing file tyhjgpo568

| Info similarity | $\frac{1}{\sqrt{6}}$ |
Extracting similarity vector

### Issue 1

<table>
<thead>
<tr>
<th>Event ID</th>
<th>Time</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>[info]</td>
<td>00:00:01.344</td>
<td>Panic 12 - unable to perform update xrtx</td>
</tr>
<tr>
<td>[crit]</td>
<td>00:00:02.232</td>
<td>[event_id 150] Restarting server xxx</td>
</tr>
<tr>
<td>[info]</td>
<td>00:00:02.345</td>
<td>[event_id 23] Start writing file xfgrew21e3456</td>
</tr>
<tr>
<td>[error]</td>
<td>00:00:03.232</td>
<td>[event_id 25] failed to write file xfgrew21e3456</td>
</tr>
<tr>
<td>[info]</td>
<td>00:00:04.252</td>
<td>[event_id 32] Panic 14 - unable to execute operation mjk</td>
</tr>
</tbody>
</table>

### Issue 2

<table>
<thead>
<tr>
<th>Event ID</th>
<th>Time</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>[info]</td>
<td>09:10:43.294</td>
<td>Panic 12 - unable to perform update gfds</td>
</tr>
<tr>
<td>[info]</td>
<td>09:10:45.594</td>
<td>[event_id 65] User 578903 authentication starts</td>
</tr>
<tr>
<td>[crit]</td>
<td>09:10:45.742</td>
<td>[event_id 68] User 578903 authentication stopped</td>
</tr>
<tr>
<td>[info]</td>
<td>00:00:02.345</td>
<td>[event_id 23] Start writing file tyhjgpo568</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Info similarity</th>
<th>Crit similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{\sqrt{6}}$</td>
<td>0</td>
</tr>
</tbody>
</table>
Extracting similarity vector

<table>
<thead>
<tr>
<th>Issue 1</th>
<th>&lt;info&gt; 00:00:01.344 [event_id 1] Panic 12 - unable to perform update xrt</th>
<th>&lt;crit&gt; 00:00:02.232 [event_id 150] Restarting server xxx</th>
<th>&lt;info&gt; 00:00:02.345 [event_id 23] Start writing file xfgrew21e3456</th>
<th>&lt;error&gt; 00:00:03.232 [event_id 25] failed to write file xfgrew21e3456</th>
<th>&lt;info&gt; 00:00:04.252 [event_id 32] Panic 14 - unable to execute operation mjk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issue 2</td>
<td>&lt;info&gt; 09:10:43.294 [event_id 1] Panic 12 - unable to perform update gfds</td>
<td>&lt;info&gt; 09:10:45.594 [event_id 65] User 578903 authentication starts</td>
<td>&lt;crit&gt; 09:10:45.742 [event_id 68] User 578903 authentication stopped</td>
<td>&lt;info&gt; 00:00:02.345 [event_id 23] Start writing file tyhjgpo568</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Info similarity</th>
<th>Crit similarity</th>
<th>Panic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/√6</td>
<td>0</td>
<td>1/√2</td>
</tr>
</tbody>
</table>
Extracting similarity vector

**Issue 1**

<iinfo> 00:00:01.344 [event_id 1] Panic 12 - unable to perform update xrty</iinfo>
<iinfo> 00:00:02.345 [event_id 23] Start writing file xfgrew21e3456</iinfo>
<iinfo> 00:00:03.232 [event_id 25] failed to write file xfgrew21e3456</iinfo>
<iinfo> 00:00:04.252 [event_id 32] Panic 14 - unable to execute operation mjk</iinfo>

**Issue 2**

<iinfo> 09:10:43.294 [event_id 1] Panic 12 - unable to perform update gfds</iinfo>
<iinfo> 09:10:45.594 [event_id 65] User 578903 authentication starts</iinfo>
<iinfo> 09:10:45.742 [event_id 68] User 578903 authentication stopped</iinfo>
<iinfo> 00:00:02.345 [event_id 23] Start writing file tyhjgpo568</iinfo>

<table>
<thead>
<tr>
<th>Info similarity</th>
<th>Crit similarity</th>
<th>Panic</th>
<th>Same Operation System</th>
<th>Time difference (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{\sqrt{6}}$</td>
<td>0</td>
<td>$\frac{1}{\sqrt{2}}$</td>
<td>True</td>
<td>12</td>
</tr>
</tbody>
</table>
Extracting similarity vector

<table>
<thead>
<tr>
<th>Info similarity</th>
<th>Crit similarity</th>
<th>Panic</th>
<th>Same Operation System</th>
<th>Time difference (days)</th>
<th>Label: Is duplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{\sqrt{6}}$</td>
<td>0</td>
<td>$\frac{1}{\sqrt{2}}$</td>
<td>True</td>
<td>12</td>
<td>True</td>
</tr>
</tbody>
</table>

**Issue 1**
<info> 00:00:01.344 [event_id 1] Panic 12 - unable to perform update xrtv
<crit> 00:00:02.232 [event_id 150] Restarting server xxx
<info> 00:00:02.345 [event_id 23] Start writing file xfgrew21e3456
<error> 00:00:03.232 [event_id 25] failed to write file xfgrew21e3456
<info> 00:00:04.252 [event_id 32] Panic 14 - unable to execute operation mjk

**Issue 2**
<info> 09:10:43.294 [event_id 1] Panic 12 - unable to perform update gfds
<info> 09:10:45.594 [event_id 65] User 578903 authentication starts
<crit> 09:10:45.742 [event_id 68] User 578903 authentication stopped
<info> 00:00:02.345 [event_id 23] Start writing file tyhgpoo568
Results
Evaluation – top [k] precision
Evaluation – users feedback

DevOps team leader: “Towards the new product release, we ran about 250-300 tests a day. Having the machine learning suggestions is one of the reasons we were able to investigate it in time.”