Research Motivation and Aim

What do we need? Extract embedded actionable information

Solution: Uncover hidden semantic structure from unstructured text

Objectives:
1. Application of Latent Dirichlet Analysis (LDA) on Google Chromium Browser ITS using metadata using:
   - Time
   - Priority
   - OS
   - Category
   - Type
   - Status
2. Multi-faceted and cross-dimensional empirical analysis of LDA output

Empirical Analysis and Results

Topic Identification
4,777 bug reports from January 2013
LDA model: 25 Topics, \( \alpha = 0.0001 \), \( \beta = 2 \)

Topic Modeling on Reopened Bugs

What are the most prevalent topics in reopened bug corpus?
Is there a link b/w re-opened bugs and their priority value?
Is there a relation between re-opened bugs and their Category value?

Topic Modeling Application

Research Methodology and Framework

Experimental Dataset

Total Bug Reports extracted = 2,15,442
Minimum Issue ID
Maximum Issue ID
Date of Reporting First Issue
Date of Reporting Last Issue
Total Number of Restricted Issues
Total Number of Comments/Bug Report
Avg. Number of Bugs Reported/Year

Unlabeled Latent Dirichlet Allocation (LDA)

Manual Label Prediction

Training Parameters:

- Number of Topics
- Dirichlet distribution parameters:
  - \( \alpha := \text{topics/document} \)
  - \( \beta := \text{words/topic} \)
  - Number of Iterations

5,502 Re-opened bugs out of which 4,988 were downloadable

Empirical Analysis and Results

Topic Identification

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top Terms</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>T16</td>
<td>autotest server packages build scripts</td>
<td>Auto-test server</td>
</tr>
<tr>
<td>T05</td>
<td>password account text sign button</td>
<td>Acc. Password text</td>
</tr>
<tr>
<td>T14</td>
<td>video flash play browser crash</td>
<td>Video crash during play</td>
</tr>
<tr>
<td>T09</td>
<td>wi-fi ethernet cellular data network</td>
<td>Network connection</td>
</tr>
<tr>
<td>T19</td>
<td>run debug libtestpy suite failed</td>
<td>Pylib debuging</td>
</tr>
</tbody>
</table>

Takeaways
Faceted browsing and analysis of bug reports leads to a more intuitive understanding and characterization of data -> better resource allocation, root-cause analysis, expertise modeling and understanding software evolution

References