Research at Scale

The Mondego Group @ UC Irvine & ISR

Crista Lopes
Pedro Martins, Assistant Project Scientist
Rohan Achar, Graduate Student
Di Yang, Graduate Student
Vainhav Saini, Graduate Student
Eugenia Gabrielova, Graduate Student
Wen Shen, Graduate Student
Farima Farmahinifarrahani, Graduate Student

ISR Research Forum, June 2017
A Couple of Projects

• Code cloning in Java, C++, Python, JavaScript (Collaboration with Jan Vitek, NEU)

• Sourcerer’s Java Build Framework

• [Is there gold in Stack Overflow data?]
  • Ask me offline!
Code Duplication
Understanding Natural Code Duplication

• Main objectives:
  • Measure it
  • Understand **what** is being cloned (qualitative analysis)
  • Understand main differences between different languages
  • Make duplication data available
# Corpus

<table>
<thead>
<tr>
<th>Counts</th>
<th>Java</th>
<th>C++</th>
<th>Python</th>
<th>JavaScript</th>
</tr>
</thead>
<tbody>
<tr>
<td># projects (total)</td>
<td>3,506,219</td>
<td>1,130,879</td>
<td>2,340,845</td>
<td>4,479,173</td>
</tr>
<tr>
<td># projects (non-fork)</td>
<td>1,859,001</td>
<td>554,008</td>
<td>1,096,246</td>
<td>2,011,875</td>
</tr>
<tr>
<td># URLs processed</td>
<td>631,390</td>
<td>554,008</td>
<td>1,096,246</td>
<td>916,059</td>
</tr>
<tr>
<td># projects (downloaded)</td>
<td>479,113</td>
<td>369,440</td>
<td>909,290</td>
<td>916,082</td>
</tr>
<tr>
<td># projects (analyzed)</td>
<td>473,562</td>
<td>364,155</td>
<td>893,197</td>
<td>903,558</td>
</tr>
<tr>
<td># files (analyzed)</td>
<td>29,592,071</td>
<td>61,647,575</td>
<td>31,602,780</td>
<td>135,712,428</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medians</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Files per project</td>
<td>11</td>
<td>11</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SLOC per file</td>
<td>42</td>
<td>55</td>
<td>46</td>
<td>28</td>
</tr>
<tr>
<td>Stars per project</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Commits per project</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>
Data Processing Pipeline

Software Projects → Tokenization → Tokens and Facts for Each File → File-hash Reduction → Group of Distinct Files → Token-hash Reduction → Distinct Source Code

Distinct Source Code → SourcererCC → Project Clones
Duplication vs. # files, # commits

Java

C++

Python

JS
Types and Amounts of Duplication

Java

- Duplicate files: 9,001,505
- Unique files: 14,312,394
- Cloned files: 8,466,685

Python

- Duplicate files: 16,432,156
- Unique files: 6,949,894
- Cloned files: 4,844,125

C++

- Duplicate files: 37,613,571
- Unique files: 11,893,435
- Cloned files: 6,596,407

JavaScript

- Duplicate files: 77,300,536
- Unique files: 5,902,360
- Cloned files: 3,944,827

- Unique files: 19,157,533
Project-Level Duplication
Current Work

• DejàVu: a Web service that returns all duplicates of a given file in GitHub
• Performance improvements to clone detection
Sourcerer’s Java Build Framework
Goal

- Automatically build ALL of GitHub Java corpus

- Today:
  - 54% non-Android
SourcererJBF

Java Projects

Custom Build Files

Round 1

Successes

Failures

Repair

Character Encoding Solver

External Dependencies Solver

Round 2

Successes

Failures

FQN to JAR

Indexing

Collection

JAR Files
SourcererJBF Effectiveness

353,709 non-Android projects

- Success: 190,727 (54%)
- Round 1: 92,482 (26%)
- Round 2: 98,245 (28%)
- Failures: 162,982 (46%)
Correlation with Project Size?
Could Own Build Scripts do Better?

189,220 out of 353,709 projects (53%)

- Maven 35%
- Ant 15%
- Gradle 3%
- No Build 47%
In 189,220 projects:
JBF: 86,926 (46%)
Own: 105,973 (56%)

In 353,709 projects:
JBF: 190,727 (54%)
Own: 105,973 (30%)
Problems with Own Builds

- Security and integrity of local build system
  - Crazy things happen!
- Unknown location of compiled code
  - Maybe jar’ed, may be moved into network, etc...
- Large variation of actions, not just compilation
  - “Success” means build script succeeded, not compilation succeeded
- Builds take much longer
  - JBF: 8 secs (median)
  - Own builds: 20 secs (median)
Improving SourcererJBF Effectiveness

Success now: 54%
Success target: 67%
Doing Research with Big Data, the Bad

• Tera-byte sized datasets
  • Difficult to handle, share
• Requires $$ hardware
  • Currently: 112-core server, 512G RAM
• Processing can take weeks
  • Mistakes are expensive
• Scientific insights don’t necessarily need big data
  • Sampling
Doing Research with Big Data, the Good

• Useful applications require the whole data
• Scale presents new engineering challenges
  • Doctoral work worthy