

Applying Text Retrieval and NLP to Software Engineering Data

2016 SATToSE hackathon

Gregorio Robles

greg@gsync.urjc.es

Universidad Rey Juan Carlos (Madrid, Spain)

Bergen, July 11th 2016





(cc) 2016 Gregorio Robles et al.

Some rights reserved. This work licensed under Creative Commons Attribution-ShareAlike License. To view a copy of full license, see <http://creativecommons.org/licenses/by-sa/3.0/> or write to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Some of the figures have been taken from the Internet Source, and author and license if known, is specified.

For those images, *fair use* applies.

Who has been at the SATToSE hackathon this summer?



Goal of the hackathon

Apply the TR and NLP techniques learned in the tutorial to software engineering data from a FLOSS project

MetricsGrimoire Toolset

- Versioning: CVSanaly
(metricsgrimoire.github.io/CVSanaly)
- Bug-tracking: Bicho
(metricsgrimoire.github.io/Bicho)
- Mailing lists: MailingListStats
(metricsgrimoire.github.io/MailingListStats)
- *et al.*

The complete suite can be accessed at
metricsgrimoire.github.io

Bitergia Datasets

- Based on the MetricsGrimoire toolset, Bitergia analyzes FLOSS projects
- Results are offered in web-based dashboards
- Data used for creating these dashboards is publicly available
- ... and will be the basis of this hackathon!

OpenStack

- FLOSS platform for cloud computing
- Over 10K developers and/from 500 companies work on it
- Heavy use of code review and modern FLOSS development techniques
- Written mainly in Python

Bitergia's OpenStack Dashboard data sources

Community 38 Subprojects All history

Project Overview / Data sources

278,392 commits 5,974 developers 105,637 tickets 143,925 mail messages

Data Source	From	To (Updated on)
Tickets (Launchpad)	2010-07-13	2016-07-08
Tickets (Storyboard)	2010-08-09	2016-07-07
Mailing Lists	2010-11-11	2016-07-08
Source Code Management (Git)	2010-05-27	2016-07-08
Source Code Review	2011-07-25	2016-07-08
IRC Messages	2010-07-08	2016-07-08
QAForums	2010-10-21	2015-12-16

Our main concern as a company is to **provide products and services as transparent as possible, generating a trustable relationship with our customers**. Therefore, if you find any inconsistencies in the data, or have any other kind of feedback, please let us know by contacting us.

Thanks for your interest and support!

Get here the [MySQL database dumps](#) with the complete datasets retrieved using the [Metrics Grimoire Tools](#).
You can also download the [JSON files \(archive\)](#) used to display the information shown on this dashboard

How this dashboard was produced

Deploying a copy of the dashboard elsewhere

Obtaining data from JSON documents

Querying the database

activity.openstack.org/dash/browser/data_sources.html

Data, data, data... (I)

- Updated daily!
- Offered in two (really three) different formats:
 - 1 (compressed) SQL (as a MySQL dump)
 - 2 JSON
 - However contains only statistics
 - 3 (For the hackathon, I transformed the data partially to) CSV

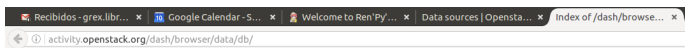
Data, data, data... (and II)

- 1 9,584 questions, 9,661 answers and 13,841 comments
- 2 108,066 tickets and 519,066 comments on them
- 3 150,963 mailing list messages (including content)
- 4 272,912 reviews and 6,838,049 comments on them
- 5 549,116 commits meta-data (including message)
- 6 22,385,190 messages on IRC
- 7 ... but no code (you can clone it from <https://github.com/openstack>)








hackathon side activity?



SQL data



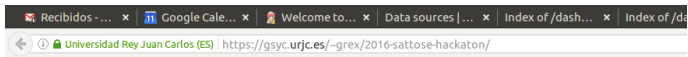
Index of /dash/browser/data/db

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 irc.mysql.7z	09-Jul-2016 17:17	262M	
 mailing_lists.mysql.7z	09-Jul-2016 16:53	81M	
 projects.mysql.7z	09-Jul-2016 17:17	7.8K	
 qaforums.mysql.7z	09-Jul-2016 17:17	4.9M	
 reviews.mysql.7z	09-Jul-2016 17:07	380M	
 source_code.mysql.7z	09-Jul-2016 16:50	52M	
 tickets.mysql.7z	09-Jul-2016 16:48	160M	












Apache/2.2.22 (Ubuntu) Server at activity.openstack.org Port 80

<http://activity.openstack.org/dash/browser/data/db/>

CSV data (I)



Index of /~grex/2016-sattose-hackaton

Name	Last modified	Size	Description
 Parent Directory		-	
 irc.csv.gz	2016-07-11 13:45	390M	
 mailing_lists.csv.gz	2016-07-11 13:46	127M	
 qaforums_answers.csv.gz	2016-07-11 13:46	2.0M	
 qaforums_comments.csv.gz	2016-07-11 13:46	1.0M	
 qaforums_questions.csv.gz	2016-07-11 13:46	4.1M	
 reviews_comments.csv.gz	2016-07-11 13:47	354M	
 reviews_issues.csv.gz	2016-07-11 13:47	8.0M	
 source_code.csv.gz	2016-07-11 13:50	61M	
 tickets_comments.csv.gz	2016-07-11 13:47	36M	
 tickets_issues.csv.gz	2016-07-11 13:47	176M	

Apache/2.4.10 (Debian) Server at gsync.urjc.es Port 443

`http://gsync.urjc.es/~grex/2016-sattose-hackathon/`

CSV data (and II)

To ease *playing* with the data and getting it known:

- First 10,000 lines of each CSV file
 - Really, header line + 9,999 lines with content
- Compressed and uncompressed
- All: all_10000s.tar.gz

http:

[//gsyc.urjc.es/~gregx/2016-sattose-hackathon/10000s](http://gsyc.urjc.es/~gregx/2016-sattose-hackathon/10000s)

What shall I do? It depends!



Possible hackathon question #1

Extraction of Domain Concepts from the Identifiers

[parse source code, extract certain identifiers, process identifiers, apply POS tagging, VSM with TFIDF. Preprocessing may be required]

(Acknowledgement goes to S. L. Abebe, P. Tonella - WCRE, 2011)

Possible hackathon question #2

Inconsistencies among the name, type, and comment of a programming entity (methods and attributes)

[parse source code, map comments to entities, split identifiers, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)]

(Acknowledgement goes to V. Arnaoudova, M. Di Penta, and G. Antoniol. - JEMSE 2015)

Possible hackathon question #3

Do developers speak/cite/talk about scientific (i.e., research) literature in their development?

[parse irc/ mailing lists/commit messages, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)...
Preprocessing may be required]

(Acknowledgement goes to Tom Mens)

Possible hackathon question #4

Do developers speak/cite/talk about architecture and/or modeling in their development?

[parse irc/ mailing lists/commit messages, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)...
Preprocessing may be required]

(Acknowledgement goes to Michel Chaudron)

Possible hackathon question #5

How do concepts (specific methodologies, etc.) expand in a project?

[parse irc/ mailing lists/ commit messages, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)...
Preprocessing may be required]

(Acknowledgement goes to me ;)...)

Go hack!

(There will be a small present for participants!)



Hack! Links where to start from

Tools

<http://metricsgrimoire.github.io/>

Code

<https://github.com/openstack>

SQL

<http://activity.openstack.org/dash/browser/data/db/>

CSV

<http://gsyc.urjc.es/~grex/2016-sattose-hackathon/>

CSV - first 10,000 lines

<http://gsyc.urjc.es/~grex/2016-sattose-hackathon/1000s>

Applying Text Retrieval and NLP to Software Engineering Data

2016 SATToSE hackathon

Gregorio Robles

greg@gsync.urjc.es

Universidad Rey Juan Carlos (Madrid, Spain)

Bergen, July 11th 2016

