Contextual Documentation Referencing on Stack Overflow

Transactions on Software Engineering
ESEC/FSE 2020 Journal First

Sebastian Baltes
@s_baltes

empirical-software.engineering

THE UNIVERSITY of ADELAIDE

Thanks to my co-authors!

Contextual Documentation Referencing on Stack Overflow

Sebastian Baltes, Christoph Treude, Martin P. Robillard

Abstract—Software engineering is knowledge-intensive and requires software developers to continually search for knowledge, often on community question answering platforms such as Stack Overflow. Such information sharing platforms do not exist in isolation, and part of the evidence that they exist in a broader software documentation ecosystem is the common presence of hyperlinks to other documentation resources found in forum posts. With the goal of helping to improve the information diffusion between Stack Overflow and other documentation resources, we conducted a study to answer the question of how and why documentation is referenced in Stack Overflow threads. We sampled and classified 759 links from two different domains, regular expressions and Android development, to qualitatively and quantitatively analyze the links' context and purpose, including attribution, awareness, and recommendations. We found that links on Stack Overflow serve a wide range of distinct purposes, ranging from citation links attributing content copied into Stack Overflow, over links clarifying concepts using Wikipedia pages, to recommendations of software components and resources for background reading. This purpose spectrum has major corollaries, including our observation that links to documentation resources are a reflection of the information needs typical to a technology domain. We contribute a framework and method to analyze the context and purpose of Stack Overflow links, a public dataset of annotated links, and a description of five major observations about linking practices on Stack Overflow. Those observations include the above-mentioned purpose spectrum, its interplay with documentation resources and applications domains, and the fact that links on Stack Overflow often lack context in form of accompanying quotes or summaries. We further point to potential tool support to enhance the information diffusion between Stack Overflow and other documentation resources.

Index Terms—Community Question Answering, Software Documentation, Information Diffusion, Hyperlinks, Stack Overflow

Software Engineering is Knowledge-intensive

OpenJDK

stackoverflow

ORACLE Java Documentation

developers

WIKIPEDIA

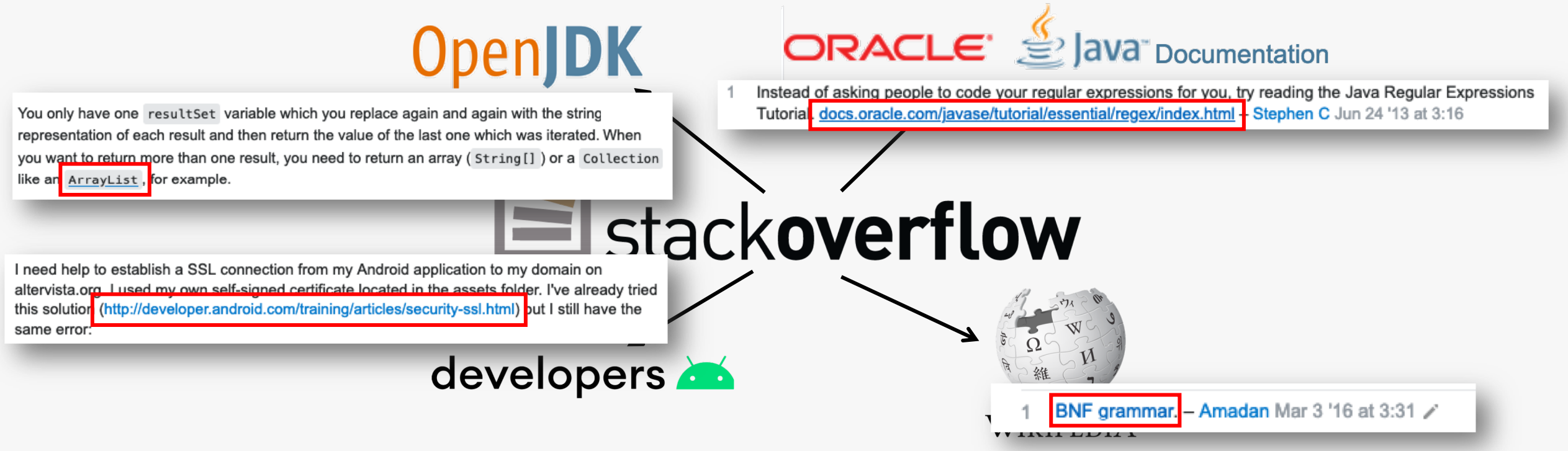
OpenJDK

Those documentation resources do not exist in **isolation**...

developers 

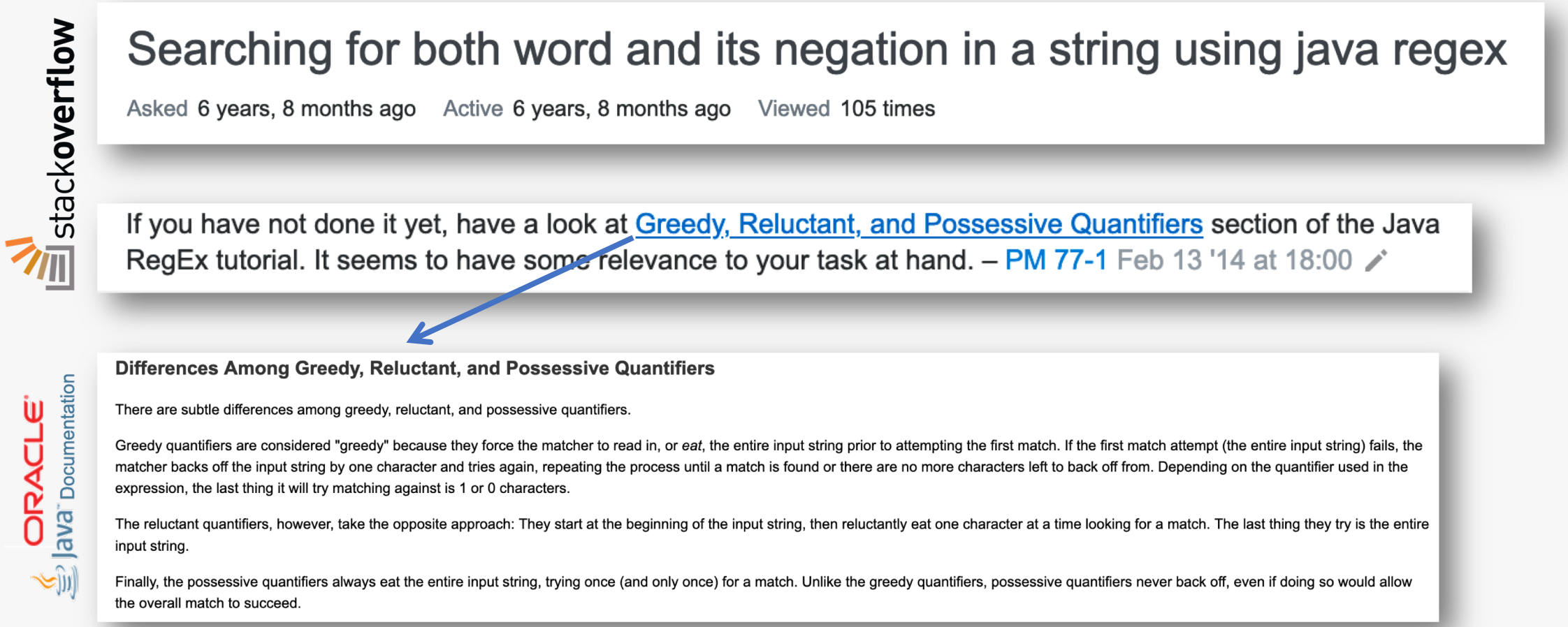
WIKIPEDIA

They form a **documentation ecosystem** with Stack Overflow as a major **information broker**



Best Case Scenario

Recommendation of specific information relevant to the thread:



Worst Case Scenario

Not adding any information while making fun of other users:

stackoverflow

java regular expression to discover regular expression

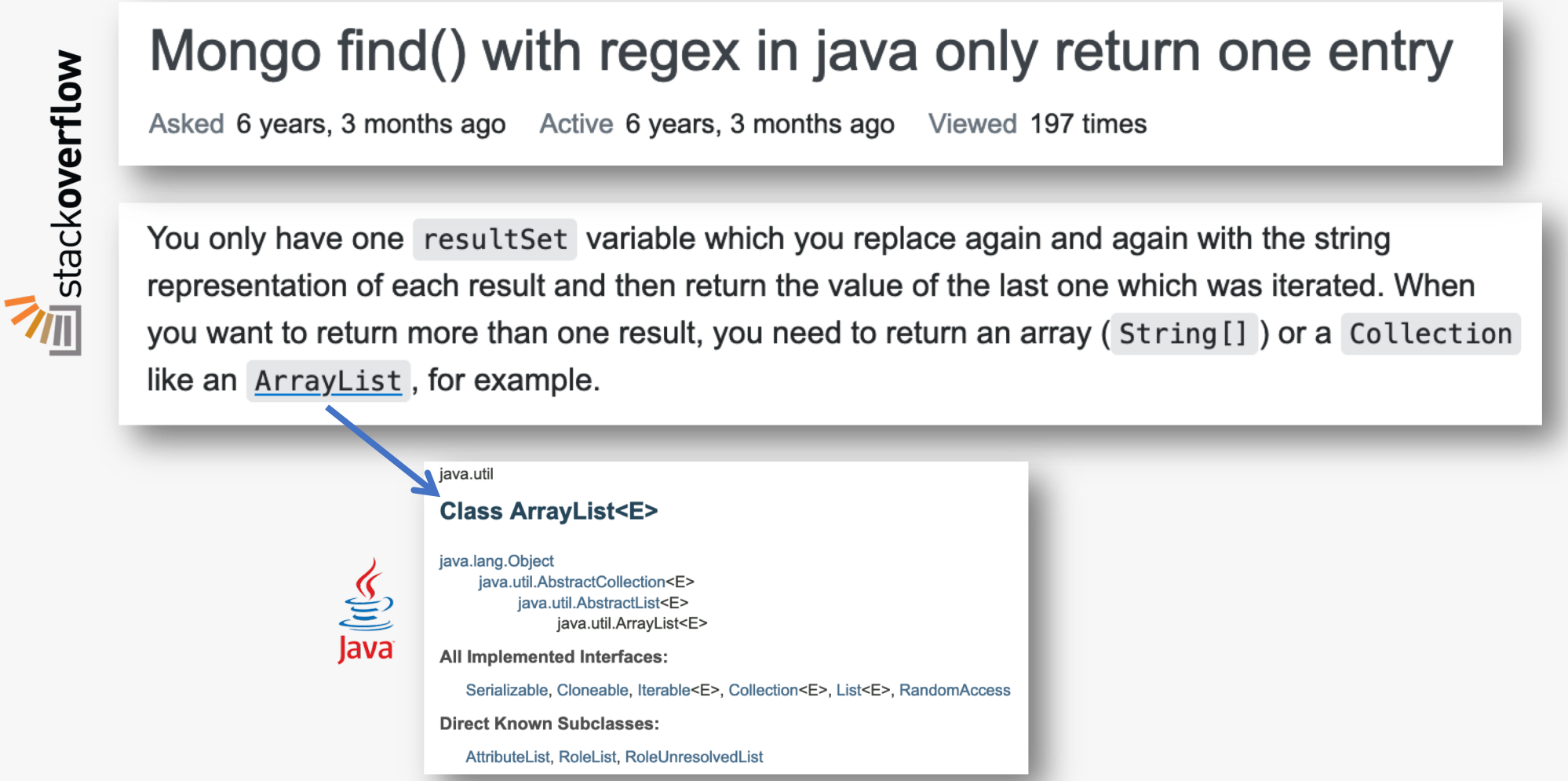
Asked 5 years, 4 months ago Active 3 years, 8 months ago Viewed 73 times

A regex to identify regexes... reminds me of this xkcd – tobias_k Jun 18 '15 at 8:56

xkcd

Linked Mentions

Frequent use case of links on Stack Overflow:

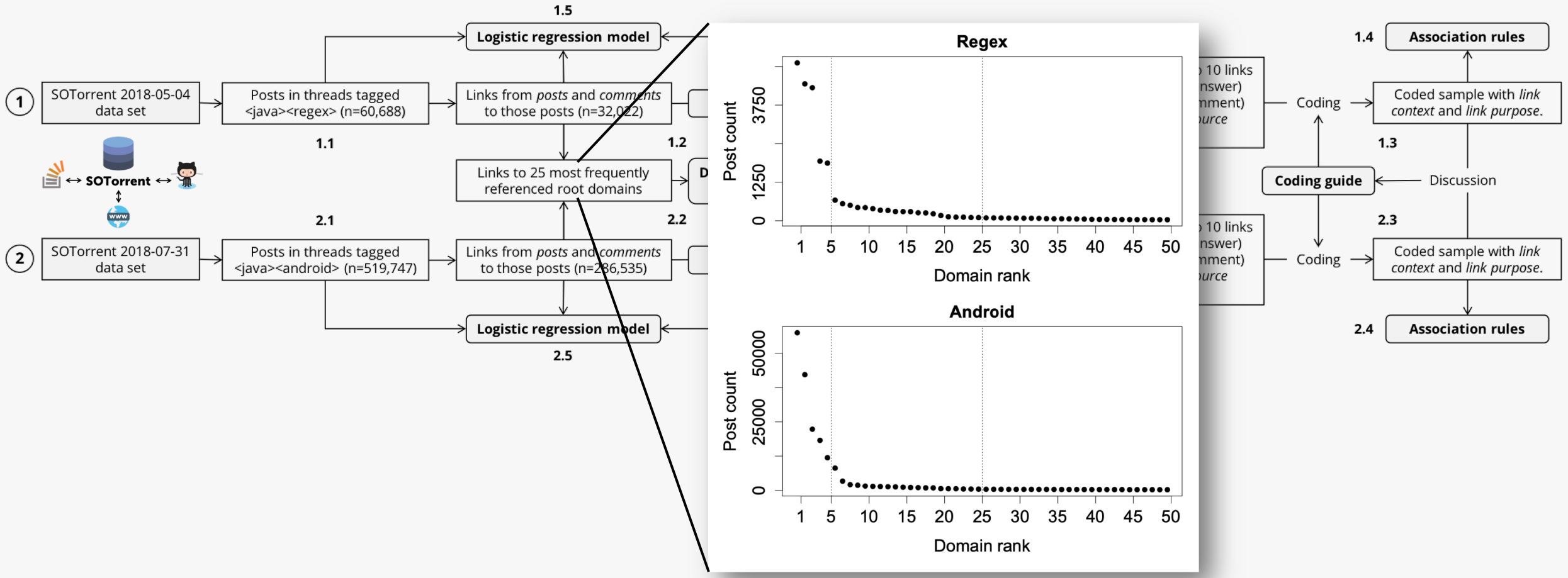


How and **why** are documentation resources referenced in Stack Overflow threads?

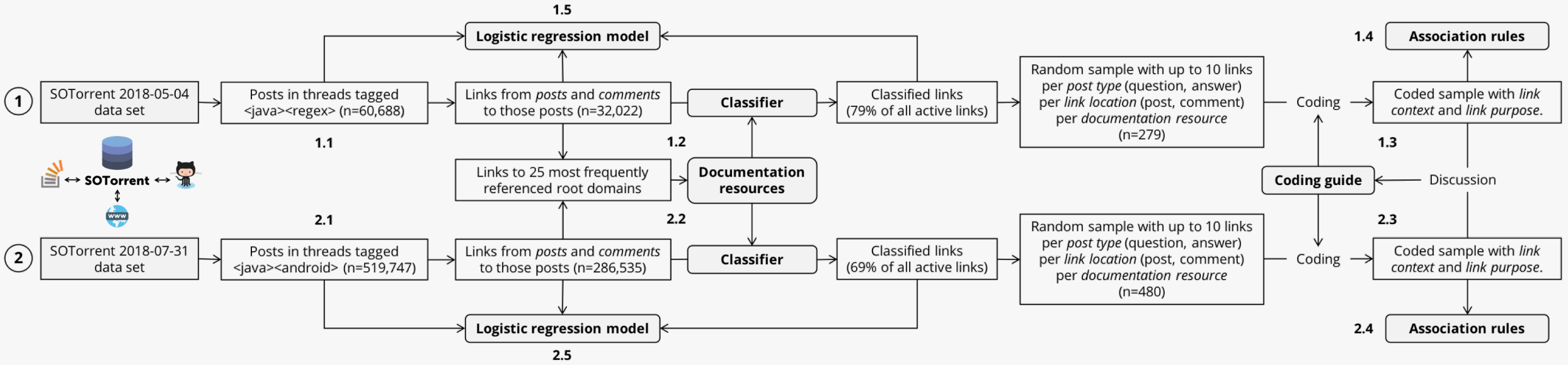
We studied:

- **Two cases…**
  (Java regular expressions and Android)
- …using a **mixed-methods** research design…
  (regex-based URL classifier, qualitative coding, association rule mining, logistic regression models)
- …focusing on link **context** (how) and link **purpose** (why).

Study Design



Study Design



Documentation Resource Categories

Resource Category	#Links in *Regex*		#Links in *Android*	
All	35,022	(100.0%)	286,535	(100.0%)
Classified	25,917	(74.0%)	185,857	(64.9%)
Documentation	**15,430**	(44.1%)	**150,630**	(52.6%)
NotDocumentation	10,487	(29.9%)	35,227	(12.3%)
NotClassified	7,115	(20.3%)	83,989	(29.3%)
InvalidOrDead	1,990	(5.7%)	16,689	(5.8%)
Documentation	**15,430**	(100.0%)	**150,630**	(100.0%)
StackOverflow	5,656	(36.7%)	64,610	(42.9%)
JavaAPI	5,093	(33.0%)	7,403	(4.9%)
IndependentTutorial	2,419	(15.7%)	6,600	(4.4%)
JavaReference	957	(6.2%)	3,860	(2.6%)
Wikipedia	787	(5.1%)	5,218	(3.5%)
OtherAPI	253	(1.6%)	644	(0.4%)
OtherReference	262	(1.7%)	6,514	(4.3%)
OtherForum	3	(0.0%)	549	(0.4%)
AndroidAPI	N/A	(0.0%)	28,690	(19.0%)
AndroidReference	N/A	(0.0%)	23,421	(15.5%)
AndroidIssue	N/A	(0.0%)	1,301	(0.9%)
YouTube	N/A	(0.0%)	1,820	(1.2%)

Classifier available:

https://github.com/sbaltes/condor

Documentation Resource Matching

Example:
Matcher for documentation resource *StackOverflow*

Domains:
```
^https?://((www|pt|ru|es)\\.)?stackoverflow\\.com
```

Paths:
```
/(a|q|questions)/[\\d]+.*
/revisions.*
/posts/\\d+/revisions.*
/posts/comments.*
```

Most Frequently Referenced Domains

<java><regex>

Domain	#Posts (%)	Resource Categories
stackoverflow.com	5,120 (23.5%)	*StackOverflow*, *NotDocumentation*
regex101.com	4,439 (20.4%)	*NotDocumentation*
oracle.com	4,316 (19.8%)	*JavaAPI*, *JavaReference*, *OtherForum*
ideone.com	1,933 (8.9%)	*NotDocumentation*
regular-expressions.info	1,868 (8.6%)	*IndependentTutorial*

<java><android>

Domain	#Posts (%)	Resource Categories
stackoverflow.com	57,461 (32.3%)	*StackOverflow*, *NotDocumentation*
android.com	42,199 (23.7%)	*AndroidAPI*, *AndroidReference*
imgur.com	22,339 (12.6%)	*NotDocumentation*
github.com	18,259 (10.3%)	*OtherReference*, *NotDocumentation*
google.com	11,924 (6.7%)	*AndroidIssue*, *AndroidReference*, *OtherReference*, *OtherForum*

Qualitative Coding of Context and Purpose

- Labeling context and purpose of links is **time-consuming**
- All **three authors** labeled the same **stratified samples** of links from both cases (n=759)
- We sampled (up to) 40 links per documentation resource
  (10 from questions, 10 from answers, 10 from question comments, 10 from answer comments)
- **Iterative development of coding guide**, tracking agreement and using discussion and majority vote to resolve disagreements

Context Codes

QUOTE

According to Glob page from Wikipedia, the wildcard `*` means:

> matches any number of any characters including none

SUMMARY

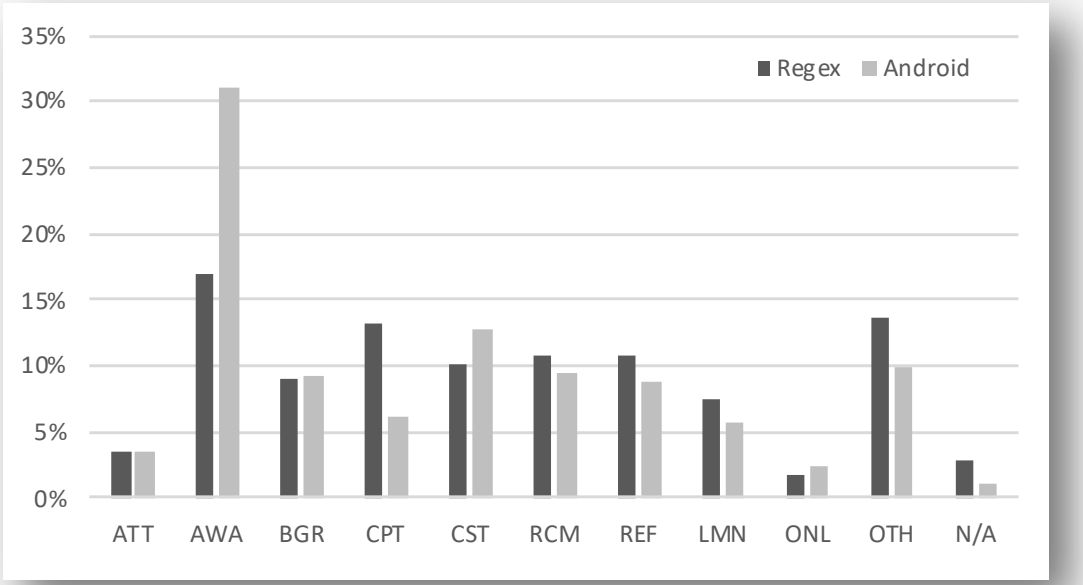
In this documentation "Communicating with Other Fragments", Google tells us that the best practice for communicating Activity and Fragment is to **implement an interface**. This interface then can be called by Fragment and to execute necessary behaviour in Activity.

LINKONLY

docs.oracle.com/javase/tutorial/java/javaOO/accesscontrol.html – Brian Roach Feb 8 '13 at 17:40

Purpose Codes

https://zenodo.org/record/2585828



ATT	ATTRIBUTION	Link to a resource simply to credit the source for material taken verbatim.
AWA	AWARENESS	Link intended to make readers aware that a certain resources exists, or provide information about the nature of its content, without necessarily endorsing it.
BGR	BACKGROUNDREADING	Link to a resource that a user thinks other users should read or watch to get better general knowledge of the topic related to the thread.
CPT	CONCEPT	Link to a resource that contains a general description of a concept that the reader should know about.
CST	CONSULTED	Link to documentation to indicate that it was consulted prior to posting.
LMN	LINKEDMENTION	Link to the element-level (class, method, field) Javadocs of an API element that is mentioned as part of the text, without more specific indication for the purpose of the link.
RCM	RECOMMENDATION	Link to resources that are landing pages for tools, libraries, API elements, or algorithms, for the purpose of recommending these.
REF	REFERENCE	Links to a resource to indicate the source of knowledge for an explicit claim, statement, or information conveyed in the post.
OTH	OTHER	Link whose purpose is other than can be captured by other codes, unclear, or unknown.

n = 759 links

Association Rule Mining

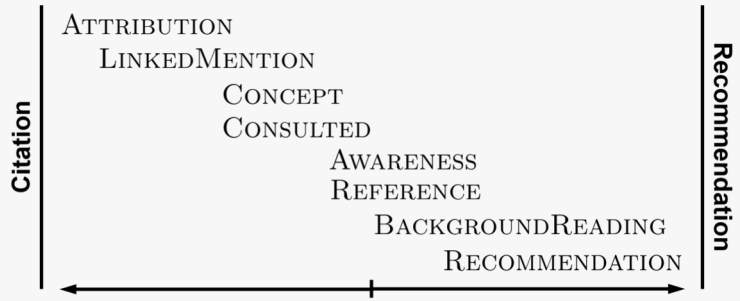
Regex

LHS	RHS	Supp	Conf	Lift	n
Wikipedia	→ CONCEPT	0.08	0.58	4.22	22
OtherAPI	→ RECOMM.	0.06	0.45	3.92	17
StackOverflow	→ AWARENESS	0.05	0.40	2.30	16
OtherReference	→ AWARENESS	0.05	0.38	2.20	13
JavaAPI	→ LINKEDMENTION	0.04	0.31	3.96	12
JavaReference	→ BACKGROUNDR.	0.04	0.30	3.24	12
Attribution	→ QUOTE	0.04	1.00	12.85	10
Reference	→ SUMMARY	0.08	0.73	3.96	22

Android

LHS	RHS	Supp	Conf	Lift	n
StackOverflow	→ AWARENESS	0.05	0.59	1.9	23
Wikipedia	→ CONCEPT	0.05	0.56	8.7	22
OtherForum	→ AWARENESS	0.05	0.55	1.8	22
AndroidIssue	→ AWARENESS	0.04	0.50	1.6	20
IndependentTut.	→ AWARENESS	0.03	0.41	1.3	16
JavaReference	→ BACKGROUNDR.	0.03	0.38	4.0	15
JavaAPI	→ RECOMM.	0.03	0.35	3.7	14
Youtube	→ AWARENESS	0.03	0.33	1.0	13
OtherReference	→ AWARENESS	0.03	0.32	1.0	12
AndroidReference	→ BACKGROUNDR.	0.02	0.28	3.0	11
AndroidReference	→ AWARENESS	0.02	0.28	0.9	11
JavaAPI	→ LINKEDMENTION	0.02	0.28	4.8	11
OtherAPI	→ RECOMM.	0.02	0.28	2.9	11
IndependentTut.	→ CONSULTED	0.02	0.26	2.0	10
AndroidAPI	→ RECOMM.	0.02	0.25	2.6	10
Attribution	→ QUOTE	0.03	0.88	18.22	15
Reference	→ SUMMARY	0.07	0.74	6.74	31

Selected Results



- **Purpose spectrum:** From citations (not necessarily meant to be consulted) to recommendations (explicit requests to follow a link).
  - Relativities Stack Overflow's recommendation to add context to every link: Adding context (summaries/quotes) important for links on the recommendation end but less important for links primarily included for citation purposes.

- **Links are reflection of the information needs** typical to a technology domain.
  - *Example:* Links to concept descriptions were twice as common in threads about regular expressions compared to Android.

- **Improving the efficiency of information diffusion** between Stack Overflow and the broader software documentation ecosystem.
  - *Example:* Developing a tool to assist readers of Stack Overflow threads by automatically classifying links along the purpose spectrum, helping them to decide whether a link is worth following.

Sebastian Baltes
@s_baltes
empirical-software.engineering