# Unlocking insights hidden in software documentation

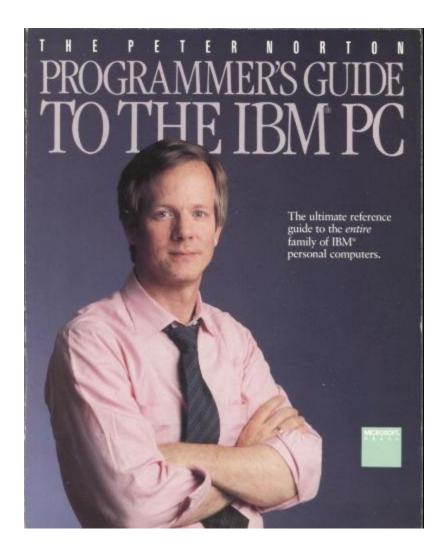
Christoph Treude

ARC DECRA Fellow and Senior Lecturer, University of Adelaide

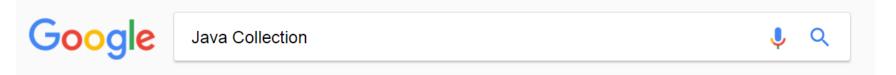


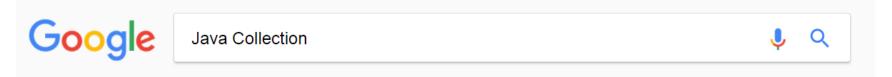


#### **Software Documentation**

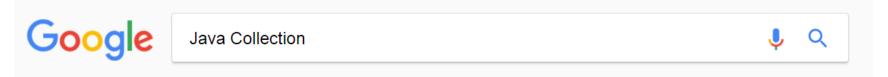


[1985]

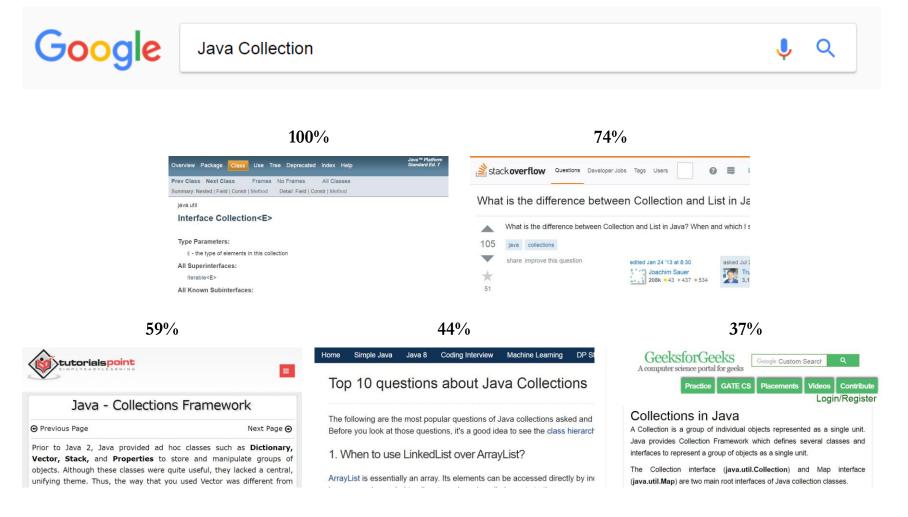












#### 162 different domains in the top 10 for 99 queries

C. Treude and M. Aniche. Where does Google find API documentation? WAPI '18: Workshop on API Usage and Evolution, p. 19-22

Tensorflow Python API: 309 different domains in the top 10 for 2,192 queries



Tensorflow Python API: 309 different domains in the top 10 for 2,192 queries



jQuery Event API: 75 different domains in the top 10 for 57 queries



C. Treude and M. Aniche. Where does Google find API documentation? WAPI '18: Workshop on API Usage and Evolution, p. 19-22

#### Navigating documentation is not trivial



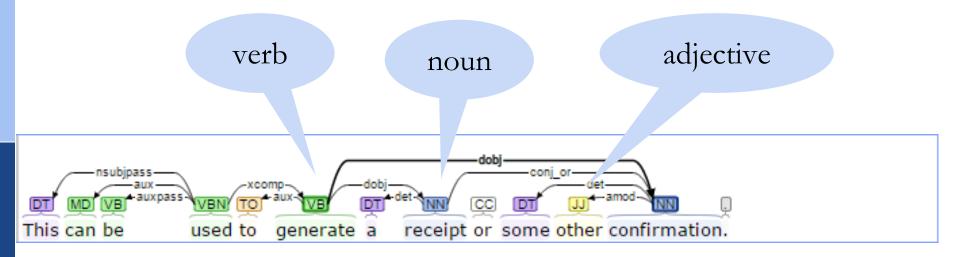
#### Navigating documentation is not trivial



#### Common Tasks



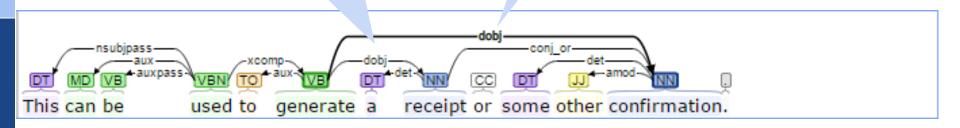
#### Extracting tasks from documentation



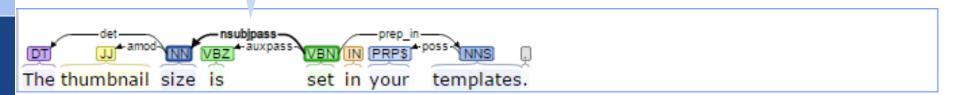
direct object:

generate
confirmation

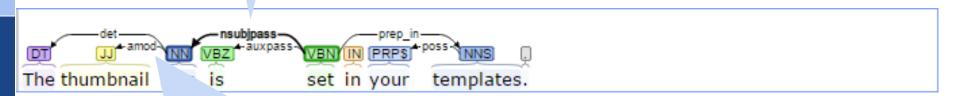
direct object:
generate receipt



passive nominal
subject: set size



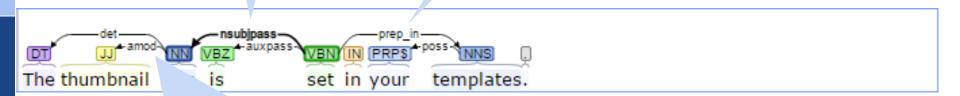
passive nominal
subject: set size



adjective modifier: set thumbnail size

preposition:set thumbnail size in templates

passive nominal
subject: set size



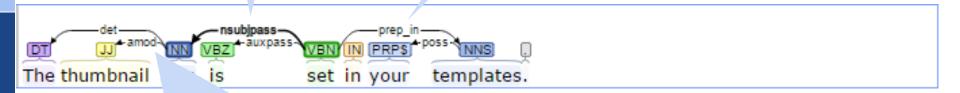
# adjective modifier: set thumbnail size

passive nominal
subject: set size

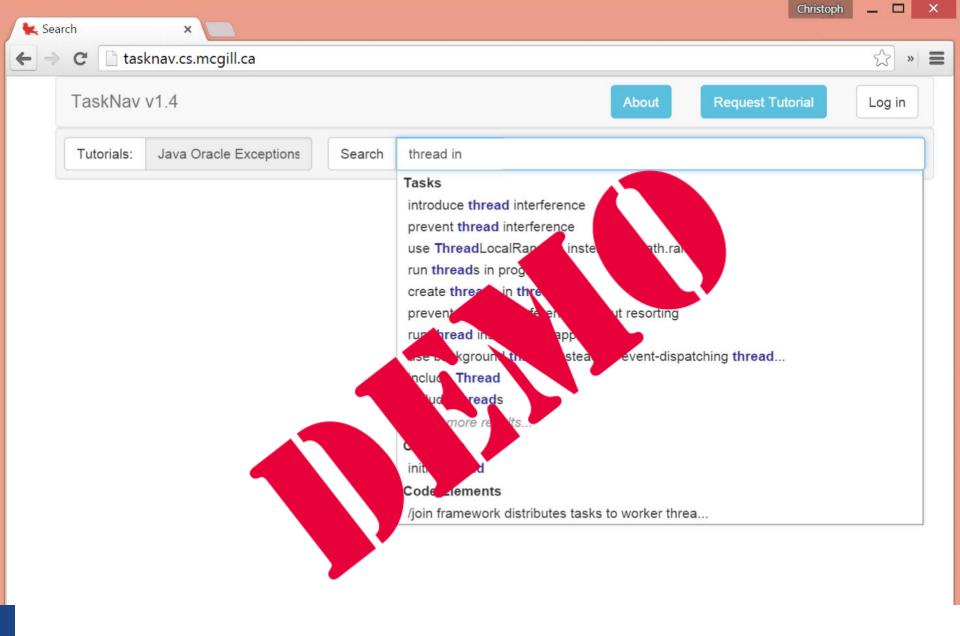
preposition:set thumbnail size in templates

#### online demo:

http://task-phrases.herokuapp.com/



## adjective modifier: set thumbnail size



C. Treude, M. Sicard, M. Klocke, and M. P. Robillard. TaskNav: Task-based Navigation of Software Documentation. ICSE '15: 37th Int'l. Conf. on Software Engineering, p. 649-652



Java Collection







59% 44% 37%

Simple Java

Coding Interview

Machine Learning



Top 10 questions about Java Collections

The following are the most popular questions of Java collections asked and Before you look at those questions, it's a good idea to see the class hierarch

1. When to use LinkedList over ArrayList?

ArrayList is essentially an array. Its elements can be accessed directly by incomparison.

Computer science portal for geeks

Practice GATE CS Placements Videos Contribute
Login/Register

Collections in Java

A Collection is a group of individual objects represented as a single unit.

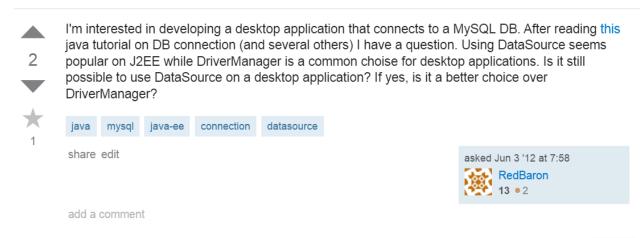
Java provides Collection Framework which defines several classes and interfaces to represent a group of objects as a single unit.

The Collection interface (java.util.Collection) and Map interface

(java.util.Map) are two main root interfaces of Java collection classes.



#### DataSource and DriverManager on J2SE





#### I'll quote the javadoc:

2 Answers

2



An alternative to the DriverManager facility, a DataSource object is the preferred means of getting a connection.

active

oldest

votes

Normally you use DriverManager when you just want a connection for one time while with DataSource you get other features such as connection pooling and distributed transactions.

Hope this helps.

add a commen

#### insight sentence

a sentence from Stack

Overflow that is related to a particular API type and that provides insight not contained in the API documentation of that type



	A	В	C		D	E
1	sentence	answer_i	type	sentence		meaningful?
816	2817	234	DriverManager	Most applications these days use the DataSource interface	and do not interact with the DriverManager like this.	у
817	32	3	AudioClip	Most media formats (sound, image, video etc.) are what is	known as 'sontainer formats'	W.
818	1654	123	String	Most random number generators are, in fact, "pseudo ran-	Coding Guide for "Meaning	
819	891	83	Serializable	My guess would be the possibility to consistently extend the	The following kinds of society of the following kinds	ningful on their own:
820	1834	149	String	My opinion, and it's only that, is that as soon as you start	The senter     Example      to another se     mutable list.	ntence.
821	3687	314	JPanel	My suggestion is to go through the Oracle Swing Tutorial		bish are red as a bod
822	1866	155	String	NB: the compare is done on 'usuario' because that's guara	wit codin σ	, which are not resolved
823	3733	319	JPanel	Neither cardsPanel1 nor cardsPanel2 (dark gray) have an	wit coding (ce a [n][3]-Matrix.	
824	286	14	Image	Next I extend JLabel:	quido	
825	286	14	JPanel	Next I extend JLabel:	guide  acing a code snippet (often indicated by a colon).  create a factory method:	
826	3408	287	ArrayList	No, ArrayList of ArrayLists cannot contain other elements		
827	3719	316	JPanel	None of these panels will be visible unless you add them t	s grammatically incomplete.	
828	2707	224	DriverManager	Normally you do not need to implement it, use the class fr		
829	2848	237	DriverManager	Normally you use DriverManager when you just want a co	dence contains communication between Stack Ov 6: Thanks to the comments I have to update my ans	
830	1000	101	Serializable	Not just Serializable.	ne sentence references code elements that come from u	ser examples rather than the
831	736	67	PropertyChange	Not likely 3, but 3.1.		
832	275	13	Image	Not really a problem when loading from the application jar		
833	275	13	JPanel	Not really a problem when loading from the application jar		
834	1650	122	String	Not too shabby.		
835	3360	279	ArrayList	Not with my solution!		
836	969	95	Serializable	Note that all non-static non-transient fields of A and B mus	The sentence contains a reference to something that is not an obvious part of the API	
837	102	7	AudioClip	Note that Class.getResource doesn't take a full file name	("block" in the example)  Example: The block size is parameterized for run-time performance optimization.	
838	1842	151	String	Note that in Java SE 7 and later you can use a String obje	The sentences starts with "but", "and", "or" etc.	
839	850	77	Serializable	Note that it still invokes defaultWriteObject for the "easy" f	Example: And then another instance of string with content Hello world.	lello world.
840	1015	105	Serializable	Note that Serializable is just a marker interface.	<ul> <li>The sentence is a generic statement that is unrelated to the API type.</li> <li>Example: Most random number generators are, in fact, pseudo random.</li> </ul>	
841	1864	155	String	Note that string constants are usually "interned" such that	The sentence contains a comparison that is incomplete (i.e., one part of the comparison is missing).	
842	3125	247	ArrayList	Note that the returned type for asList() is a List using a co		
843	1575	112	String	Note that the second parameter (i.e. Charsets.	Example: char[] is less vulnerable.	
844	466	39	Image	Note that you can find the background color component by	<ul> <li>The sentence resulted from a parsing error.</li> <li>Example: Because the java.io.</li> </ul>	

C. Treude and M. P. Robillard. Augmenting API Documentation with Insights from Stack Overflow. ICSE '16: 38th Int'l. Conference on Software Engineering, p. 392-403

• The sentence contains an explicit reference to a piece of context that's missing. Example: What you're trying to do here is rather unusual, to say the least.

Other sentences are considered meaningful

#### Supervised Insight Sentence Extractor

1	cosine similarity between sentence and most similar
	sentence in API documentation

- 2 average cosine similarity between sentence and all sentences in API documentation
- 3 question user acceptance rate
- 4 answer age
- 5 answer time difference to question
- 6 question score
- 7 question favorites
- 8 question user reputation
- 9 question views
- 10 number of nouns followed by verb in present tense, third person singular in sentence

- 11 question age
- sentence starts with noun followed by verb in present tense, third person singular
- 13 number of tokens in sentence
- 14 position of API element in sentence
- 15 number of occurrences of API element in sentence
- 16 answer score
- 17 answer size
- 18 number of nouns in sentence
- 19 sentence starts with noun
- 20 number of characters that are code
- 21 number of occurrences of the verb "be" in sentence

#### Supervised Insight Sentence Extractor

- 1 cosine similarity between sentence and most similar sentence in API documentation
- 2 average cosine similarity between sentence and all sentences in API documentation
- 3 question user acceptance rate
- 4 answer age
- 5 answer time difference to question
- 6 question score
- 7 question favorites
- 8 question user reputation
- 9 question views
- 10 number of nouns followed by verb in present tense, third person singular in sentence

- 11 question age
- sentence starts with noun followed by verb in present tense, third person singular
- 13 number of tokens in sentence
- 14 position of API element in sentence
- 15 number of occurrences of API element in sentence
- 16 answer score
- 17 answer size
- 18 number of nouns in sentence
- 19 sentence starts with noun
- 20 number of characters that are code
- 21 number of occurrences of the verb "be" in sentence

### Supervised Insight Sentence Extractor

- 1 cosine similarity between sentence and most similar sentence in API documentation
- 2 average cosine similarity between sentence and all sentences in API documentation
- 3 question user acceptance rate
- 4 answer age
- 5 answer time difference to question
- 6 question score
- 7 question favorites
- 8 question user reputation
- 9 question views
- 10 number of nouns followed by verb in present tense, third person singular in sentence

- 11 question age
- sentence starts with noun followed by verb in present tense, third person singular
- 13 number of tokens in sentence
- 14 position of API element in sentence
- 15 number of occurrences of API element in sentence
- 16 answer score
- 17 answer size
- 18 number of nouns in sentence
- 19 sentence starts with noun
- 20 number of characters that are code
- 21 number of occurrences of the verb "be" in sentence

#### Comparative Study

Edit this form

### Augmenting Java API Documentation with Sentences from Stack Overflow

\* Required

#### java.io.InputStream

http://docs.oracle.com/javase/7/docs/api/java/io/InputStream.html

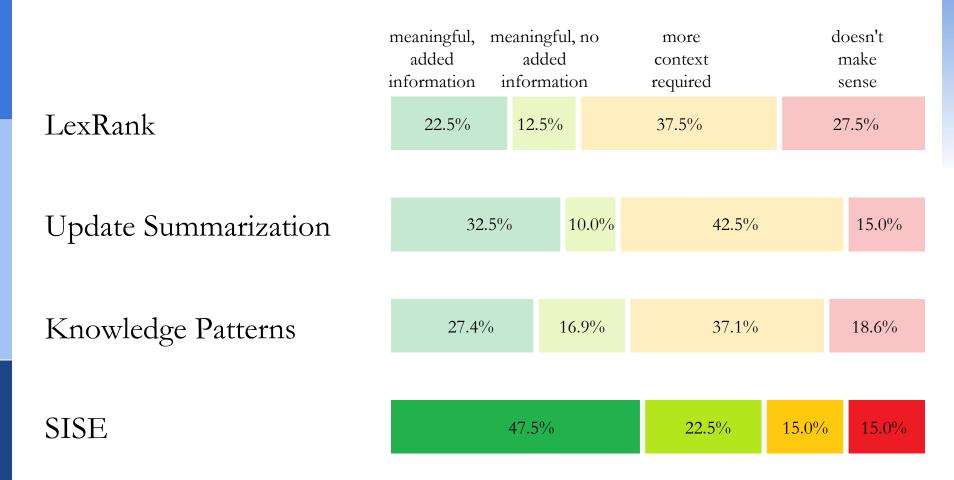
In the following, we are showing you sentences from Stack Overflow that are related to Java's InputStream. They were automatically extracted using different approaches which we are comparing in this study. Please indicate for each sentence whether it is meaningful and whether it adds useful information. The answer options are the same for each sentence.

The best practice is to put your entire code into a try{ ... }catch block and add a finally block where

#### FileInputStream is an inputStream. \*

- The sentence is meaningful and adds useful information not found in the API documentation.
- The sentence is meaningful, but does not add useful information to the API documentation.
- The sentence requires more context to understand.
- The sentence does not make sense to me.

### **Comparative Study**



C. Treude and M. P. Robillard. Augmenting API Documentation with Insights from Stack Overflow. ICSE '16: 38th Int'l. Conference on Software Engineering, p. 392-403



Java Collection







59% 44% 37%

Simple Java

Coding Interview

Machine Learning



Top 10 questions about Java Collections

The following are the most popular questions of Java collections asked and Before you look at those questions, it's a good idea to see the class hierarch

1. When to use LinkedList over ArrayList?

ArrayList is essentially an array. Its elements can be accessed directly by incomparison.

Computer science portal for geeks

Practice GATE CS Placements Videos Contribute
Login/Register

Collections in Java

A Collection is a group of individual objects represented as a single unit.

Java provides Collection Framework which defines several classes and interfaces to represent a group of objects as a single unit.

The Collection interface (java.util.Collection) and Map interface

(java.util.Map) are two main root interfaces of Java collection classes.

javascript × 1672211

JavaScript (not to be confused with Java) is a high-level, dynamic, multi-paradigm, weakly-typed language used for both client-

763 asked today, 5021 this week

java × 1451316

Java (not to be confused with JavaScript or JScript or JS) is a general-purpose object-oriented programming language designed to

523 asked today, 3312 this week

c# × 1237935

a high level, object-oriented programming language that is designed for building a variety of applications that run on the .NET

436 asked today, 2607 this week

android × 1129817

Google's mobile operating system, used for programming or developing digital devices (Smartphones, Tablets, Automobiles, TVs,

476 asked today, 2762 this week

python × 1012032

a dynamic, strongly typed, object-oriented, multipurpose programming language, designed to be quick (to learn, to use, and to

776 asked today, 4942 this week

jquery × 924833

a popular cross-browser JavaScript library that facilitates Document Object Model (DOM) traversal, event handling, animations,

203 asked today, 1261 this week

c++ × 582841

a general-purpose programming language. It was originally designed as an extension to C, and keeps a similar syntax, but is now a

169 asked today, 1112 this week

ios × 574313

the mobile operating system running on the Apple iPhone, iPod touch, and iPad. Use this tag [ios] for questions related to

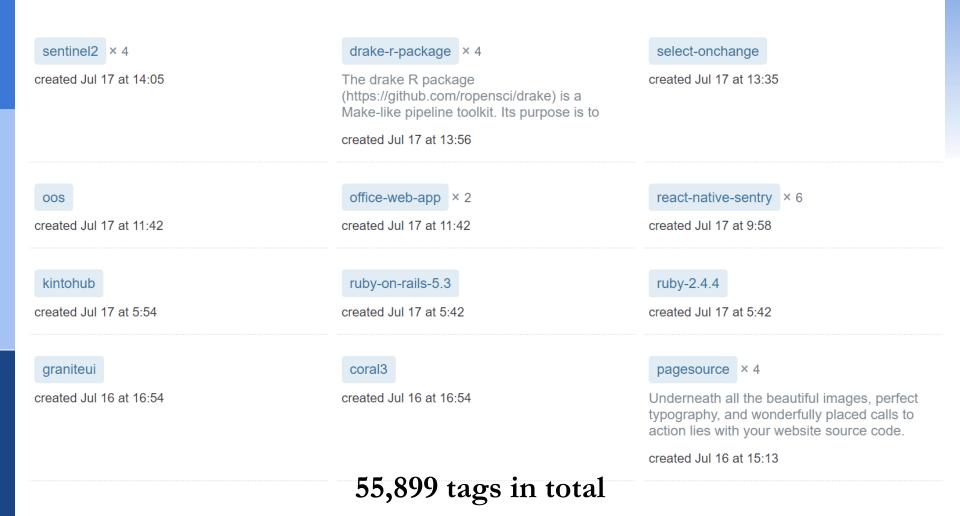
177 asked today, 1139 this week

css × 550605

a representation style sheet language used for describing the look and formatting of HTML (Hyper Text Markup Language), XML

207 asked today, 1328 this week

#### 55,899 tags in total



sentinel2 × 4

created Jul 17 at 14:05

drake-r-package × 4

The drake R package (https://github.com/ropensci/drake) is a Make-like pipeline toolkit. Its purpose is to

created Jul 17 at 13:56

select-onchange

created Jul 17 at 13:35

created Jul 16 at 15:13

#### Hypernym discovery:

Discovery of the more general concept in a is-a relationship

created Jul 17 at 5:54 created Jul 17 at 5:42 created Jul 17 at 5:42

graniteui coral3 pagesource × 4

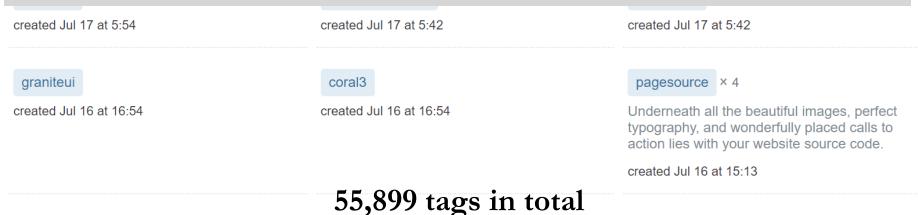
created Jul 16 at 16:54 Underneath all the beautiful images, perfect typography, and wonderfully placed calls to action lies with your website source code.

55,899 tags in total



#### Hypernym discovery:

Discovery of the more general concept in a is-a relationship







**html** markup language

**python** imperative and functional programming language

objective-c general-purpose programming languageobjective-c object oriented programming language

ruby general-purpose open-source dynamic object oriented reflective programming

language

java open-source, dynamic, reflective, object oriented, general-purpose

programming language

java class-based, object oriented, strongly typed, reflective language and run-time

environment

silverlight run time environment and multimedia framework

django open source web application framework

**xcode** integrated development environment

phpstorm commercial ide for php

**phpstorm** commercial php integrated development environment





**html** markup language

python imperative and functional programming language

Category

objective-c general-purpose programming language

objective-c object oriented programming language

ruby general-purpose open-source dynamic object oriented reflective programming

language

java open-source, dynamic, reflective, object oriented, general-purpose

programming language

java class-based, object oriented, strongly typed, reflective language and run-time

environment

silverlight run time environment and multimedia framework

django open source web application framework

**xcode** integrated development environment

phpstorm commercial ide for php

**phpstorm** commercial php integrated development environment





**html** markup language

**python** imperative and functional programming language

objective-c general-purpose programming language

objective-c object oriented programming language

ruby general-purpose open-source dynamic object oriented reflective programming

language

java open-source, dynamic, reflective, object oriented, general-purpose

programming language

java class-based, object oriented, strongly typed, reflective language and run-time

environment

silverlight run time environment and multimedia framework

django open source web application framework

**xcode** integrated development environment

phpstorm commercial ide for php

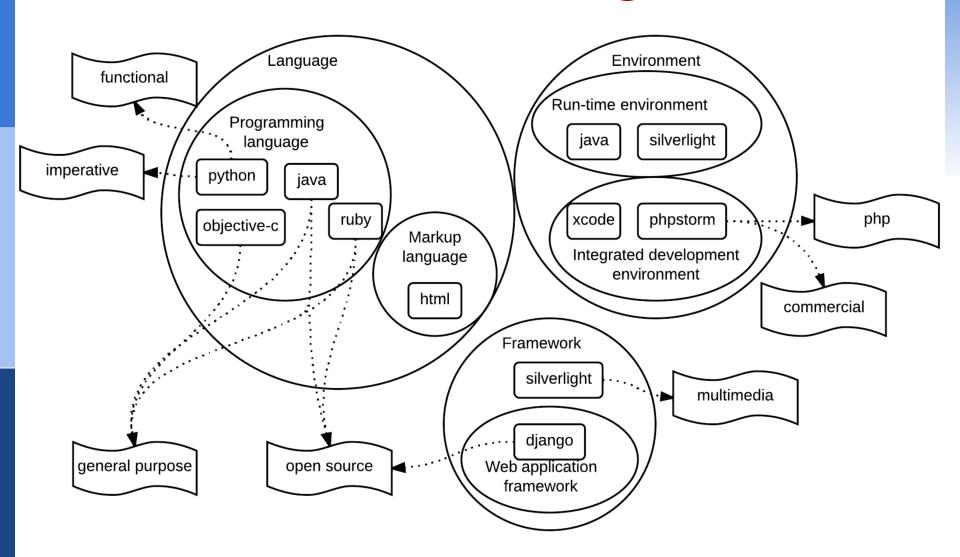
**phpstorm** commercial php integrated development environment





Attribute

#### Attributes and Categories

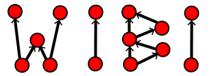


M. Nassif, C. Treude, and M. P. Robillard. Automatically Categorizing Software Technologies. To appear in IEEE Transactions on Software Engineering, 2018

## **Evaluation**

Witt





WeblsA Database



Targeted Hypernym Discovery



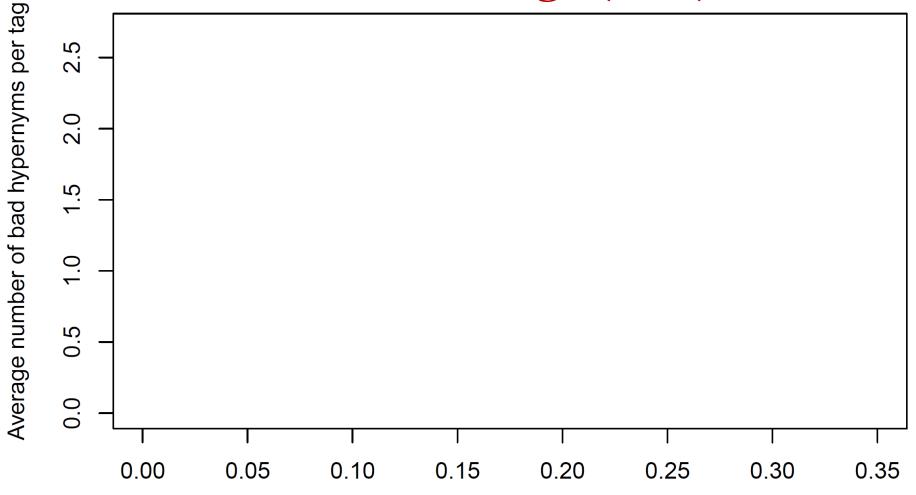
M. Nassif, C. Treude, and M. P. Robillard. Automatically Categorizing Software Technologies. To appear in IEEE Transactions on Software Engineering, 2018

## **Evaluation**

	А	В	С
1	Tag	Hypernym	Correct hypernym in the sense of software technology: Y/N
2	google-chrome-extension	Software add-ons	у
3	windows	gap	<sup>n</sup> Witt
4	midori	stellar renowned artist	n WILL
5	vector	virus	n
6	ruby-on-rails	open source full-stack web application framework	WordNet
7	normalizing	harden	n
8	root-node	widely-used abstract data-type data-structure	n A Lexical Database for English
9	windows-phone-7	Cloud clients	n
10	bloodhound	large tracker dog	n 🕊 🕊 🛼 🕊
11	reporting-services	Goods	
12	x12	major edus edus standard	n M
13	i18n-gem	Britpop musicians	n
14	ancestor	Morehead City, North Carolina	n
15	install-sequence	Installation software	n WeblsA Database
16	array-initialize	Arrangement	n
17	camera	device	у
18	sqlx	great tool	<sup>n</sup> Google
19	orm	computer programming	n Joole
20	reference	Reference	n
21	wordpress-plugin	framework	n
22	n2o	Industrial hygiene	n
23	рр	Pascal	n Targeted Hypernym Discovery
24	bing	Artificial hills	n
25	syncdb	system	n Pagant for any to of bit or specific for any to or any to order
26	ora-20000	People from Pristina	n DD Colin Company
27	active-directory	Organisation	n DBoedia Sportions

M. Nassif, C. Treude, and M. P. Robillard. Automatically Categorizing Software Technologies. To appear in IEEE Transactions on Software Engineering, 2018

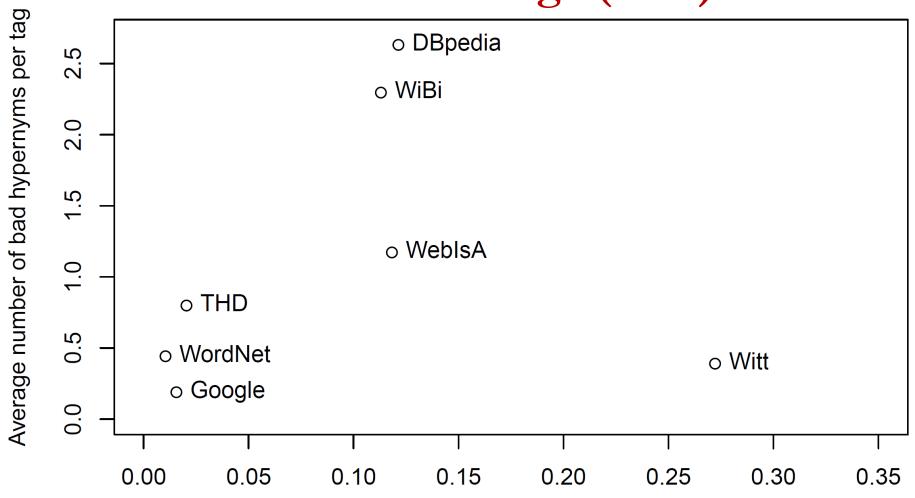
# What is this thing? (Witt)



Proportion of tags with at least one good hypernym

M. Nassif, C. Treude, and M. P. Robillard. Automatically Categorizing Software Technologies. To appear in IEEE Transactions on Software Engineering, 2018

# What is this thing? (Witt)



Proportion of tags with at least one good hypernym

M. Nassif, C. Treude, and M. P. Robillard. Automatically Categorizing Software Technologies. To appear in IEEE Transactions on Software Engineering, 2018

## Software Documentation is everywhere



Java Collection







59% 44% 37%

Simple Java

Coding Interview

Machine Learning



Top 10 questions about Java Collections

The following are the most popular questions of Java collections asked and Before you look at those questions, it's a good idea to see the class hierarch

1. When to use LinkedList over ArrayList?

ArrayList is essentially an array. Its elements can be accessed directly by incomparison.

Computer science portal for geeks

Practice GATE CS Placements Videos Contribute
Login/Register

Collections in Java

A Collection is a group of individual objects represented as a single unit.

Java provides Collection Framework which defines several classes and interfaces to represent a group of objects as a single unit.

The Collection interface (java.util.Collection) and Map interface

(java.util.Map) are two main root interfaces of Java collection classes.

# Challenges for Newcomers



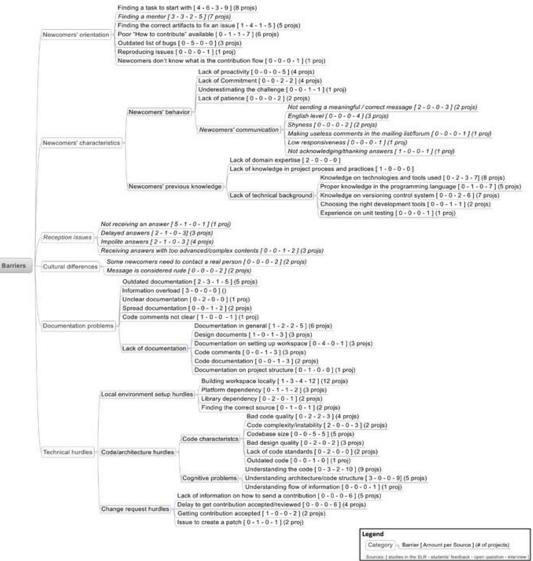
Java Collection





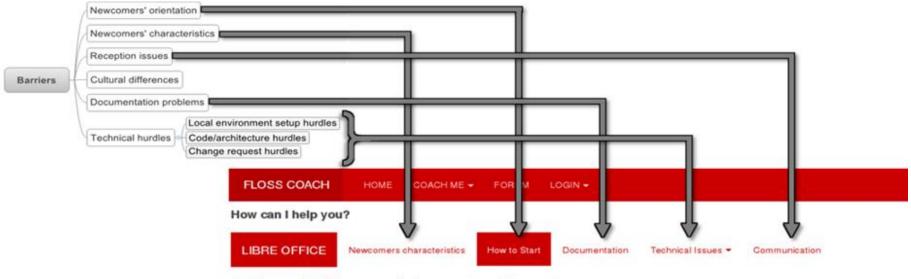


## Barriers for Newcomers to OSS projects

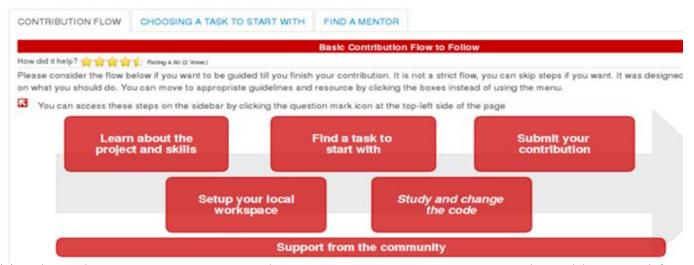


I. Steinmacher, T. U. Conte, M. A. Gerosa, and D. Redmiles. Social barriers faced by newcomers placing their first contribution in open source software projects. CSCW '15: 18th Conf. on Comp. supported coop. work & social computing, p. 1379-1392

## **FLOSSCoach**



#### LibreOffice - How to Start

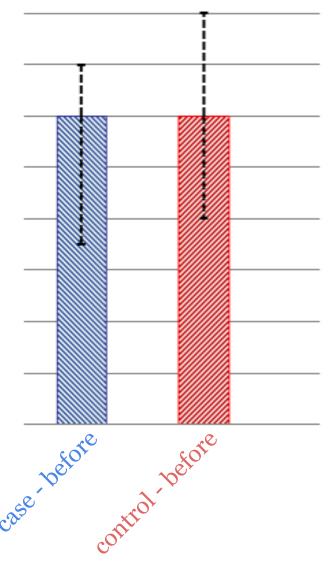


I. Steinmacher, T. U. Conte, C. Treude, and M. A. Gerosa. Overcoming Open Source Project Entry Barriers with a Portal for Newcomers. ICSE '16: 38th Int'l. Conf. on Software Engineering, p. 273-284

## **FLOSSCoach**

Self-efficacy:

"I feel comfortable with the process of contributing to an Open Source project"

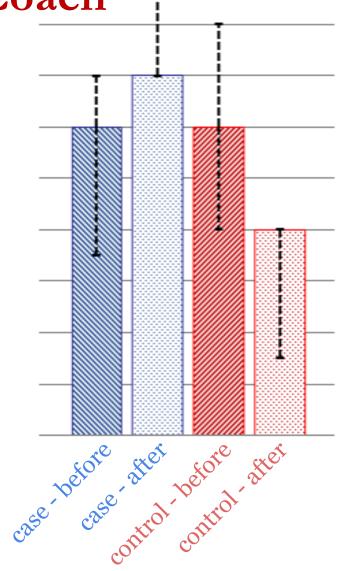


I. Steinmacher, T. U. Conte, C. Treude, and M. A. Gerosa. Overcoming Open Source Project Entry Barriers with a Portal for Newcomers. ICSE '16: 38th Int'l. Conf. on Software Engineering, p. 273-284

## **FLOSSCoach**

Self-efficacy:

"I feel comfortable with the process of contributing to an Open Source project"



I. Steinmacher, T. U. Conte, C. Treude, and M. A. Gerosa. Overcoming Open Source Project Entry Barriers with a Portal for Newcomers. ICSE '16: 38th Int'l. Conf. on Software Engineering, p. 273-284

# Söftware

**ABOUT** 

**HISTORY** 

**BACK ISSUES** 

WRITE FOR US

**SUBSCRIBE** 

**SE-RADIO** 

**BLOG** 

JOIN

# 20 Years of Open Source— Impact on Software **Engineering Practice: Call for Papers**

JULY 25, 2018

CALLS FOR PAPERS 4 MIN READ

Submission deadline: 1 Apr. 2019

Publication date: Nov./Dec. 2019











## Challenges in Analyzing Documentation



**VS** 



F. N. A. Al Omran and C. Treude. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. MSR '17: 14th Int'l. Conf. on Mining Software Repositories, p. 187-197

## Comparing NLP libraries









Only between 60% and 71% of tokens from Stack Overflow, GitHub, and the Java API Documentation were assigned the same partof-speech tag by all four libraries.

1. different tokenization

2. general part of speech

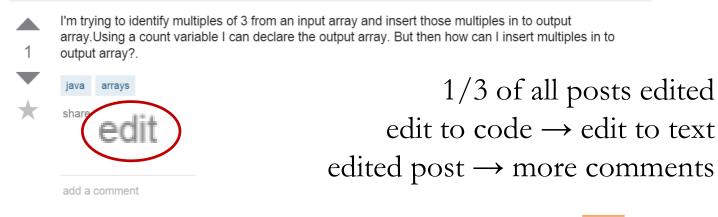
3. specific part of speech

F. N. A. Al Omran and C. Treude. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. MSR '17: 14th Int'l. Conf. on Mining Software Repositories, p. 187-197





S. Baltes, L. Dumani, C. Treude, and S. Diehl. SOTorrent: Reconstructing and Analyzing the Evolution of Stack Overflow Posts. MSR '18: 15th Int'l. Conf. on Mining Software Repositories, p. 319-330





S. Baltes, L. Dumani, C. Treude, and S. Diehl. SOTorrent: Reconstructing and Analyzing the Evolution of Stack Overflow Posts. MSR '18: 15th Int'l. Conf. on Mining Software Repositories, p. 319-330

#### Insert elements from one array to another array in Java



I'm trying to identify multiples of 3 from an input array and insert those multiples in to output array. Using a count variable I can declare the output array. But then how can I insert multiples in to output array?





1/3 of all posts edited edit to code → edit to text edited post → more comments

# MSR Challenge 2019: SOTorrent



S. Baltes, L. Dumani, C. Treude, and S. Diehl. SOTorrent: Reconstructing and Analyzing the Evolution of Stack Overflow Posts. MSR '18: 15th Int'l. Conf. on Mining Software Repositories, p. 319-330



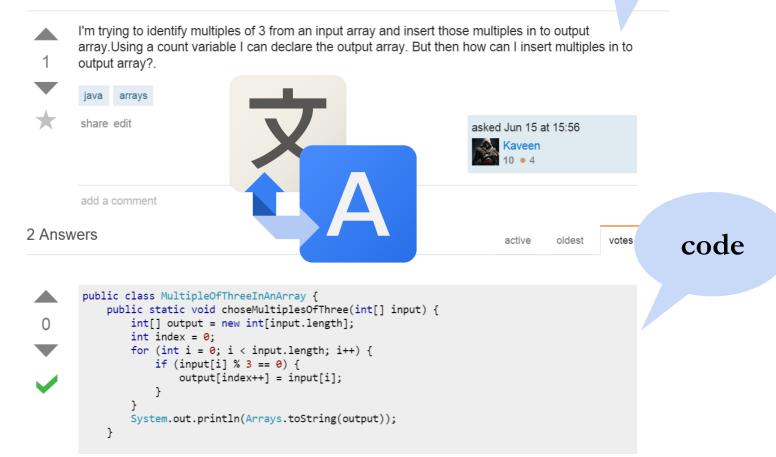
tasks Insert elements from one array to another array in Java I'm trying to identify multiples of 3 from an input array and insert those multiples in to output array. Using a count variable I can declare the output array. But then how can I insert multiples in to output array?. share edit asked Jun 15 at 15:56 Kaveen add a comment 2 Answers code active oldest votes public class MultipleOfThreeInAnArray { public static void choseMultiplesOfThree(int[] input) { int[] output = new int[input.length]; int index = 0; for (int i = 0; i < input.length; i++) {</pre> if (input[i] % 3 == 0) { output[index++] = input[i];

B. A. Campbell and C. Treude. NLP2Code: Code Snippet Content Assist via Natural Language Tasks. ICSME '17: 33rd Int'l. Conf. on Software Maintenance and Evolution, p. 628-632

System.out.println(Arrays.toString(output));

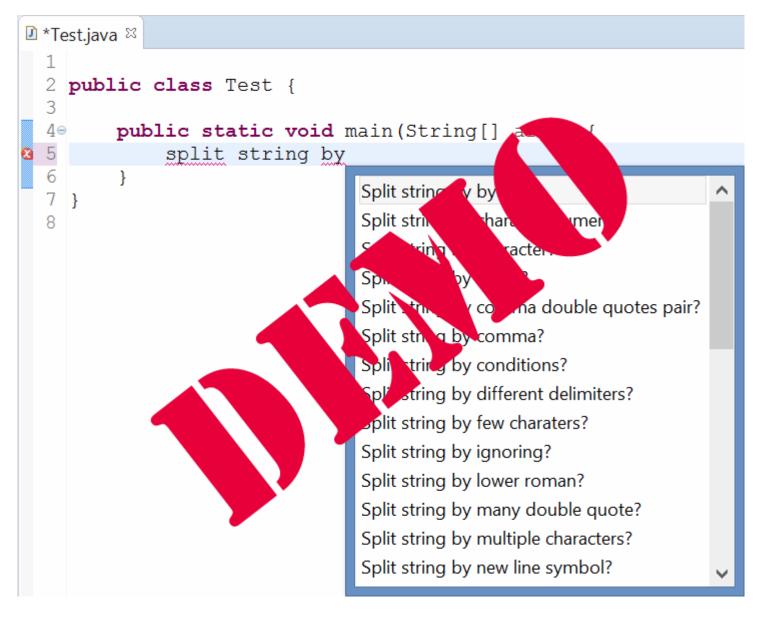
## Code Snippet Content Assist

tasks



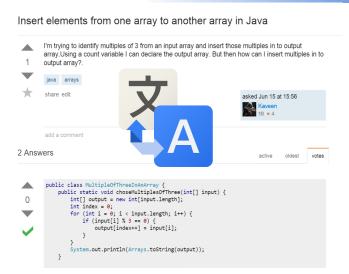
B. A. Campbell and C. Treude. NLP2Code: Code Snippet Content Assist via Natural Language Tasks. ICSME '17: 33rd Int'l. Conf. on Software Maintenance and Evolution, p. 628-632

#### https://github.com/ctreude/nlp2code



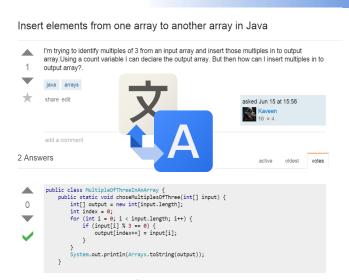
B. A. Campbell and C. Treude. NLP2Code: Code Snippet Content Assist via Natural Language Tasks. ICSME '17: 33rd Int'l. Conf. on Software Maintenance and Evolution, p. 628-632



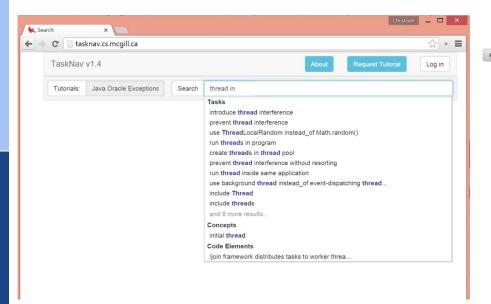


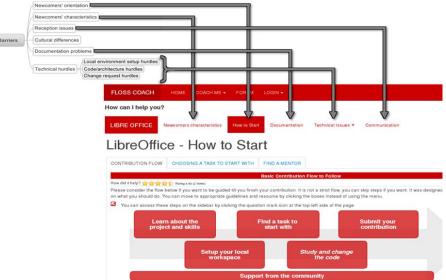
Most information needed by developers already exists somewhere,





Most information needed by developers already exists somewhere,





we're working on getting it to them where and when they need it.

# Acknowledgements



Bart Dagenais, Resulto, Canada



Gregorio Robles, URJC, Spain



Martin Robillard, McGill, Canada



Paul Adams, Zalando, Germany



Brock Campbell, U Adelaide, Australia



Igor Steinmacher, NAU, USA



Mathieu Nassif, McGill, Canada



Sebastian Baltes, U Trier, Germany



Chris Parnin, NCSU, USA



Marc Klocke, McGill, Canada



Mathieu Sicard, McGill, Canada



Stephan Diehl, U Trier, Germany



David Redmiles, UC Irvine, USA



Marco Gerosa, NAU, United States



Mauricio Aniche, TU Delft, Netherlands

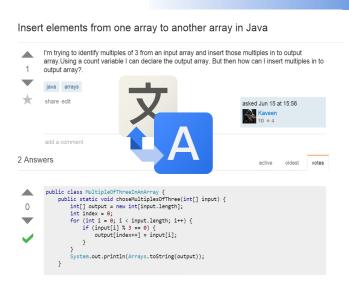


Tayana Conte, UFAM, Brazil



Fouad Al-Omran, U Adelaide, Australia





Most information needed by developers already exists somewhere,



we're working on getting it to them where and when they need it.