

# **TUTORIAL: EVALUATION OF SIMILARITIES AMONG OPEN SOURCE SOFTWARE**

**Phuong Nguyen, Juri Di Rocco, Riccardo Rubei, Davide Di Ruscio**

*Software Engineering and Architecture (SEA) Group*

*Department of Information Engineering, Computer Science and Mathematics*

*Università degli Studi dell'Aquila*

*Via Vetoio 2 -- 67100 L'Aquila, Italy*

**Email: {[phuong.nguyen](mailto:phuong.nguyen@univaq.it), [juri.dirocco](mailto:juri.dirocco@univaq.it), [riccardo.rubei](mailto:riccardo.rubei@univaq.it), [davide.diruscio](mailto:davide.diruscio@univaq.it)}@univaq.it**

We would like to thank you for your participation in our evaluation. This user study is conducted to evaluate the performance of 4 different tools for finding similar software projects, namely: **MUDABlue**, **CLAN**, **RepoPal** and **CrossSim** (proposed by the UDA team). In this document, we are going to present you a quick overview on how to properly perform the human evaluation.

A query project is the one that needs recommendation which contains a list of projects being similar to it. Given a query project, by using a similarity tool, we get a ranked list of 5 retrieved projects. We would like to know that if the 5 retrieved projects are really **similar to the query project** according to human perception. To this end, we kindly ask you to give a score to each pair of **<query project, retrieved project>** using the following guidelines.

## **1. Introduction**

We consider two open source software (OSS) projects similar if they implement similar features, or they are described by the same abstraction. Two OSS projects are similar if there is significant overlap in the requirements and functionalities of the software. In other words, we evaluate the similarity between two projects solely with regards to their functionalities, irregardless of their implementation.

There are the following examples:

- If two OSS projects implement cryptographic services to protect information then they are similar to a certain degree, even though they may have other different functionalities for different domains.
- Two text editors that are implemented by different programmers, but share many features: copy and paste, undo and redo, saving data in files using standard formats, are similar.
- An OSS project  $p_1$  that performs the sending of files across a TCP/IP network is somehow similar to an OSS project  $p_2$  that exchanges text messages between two users, i.e.  $\text{Score}(p_1, p_2)=3$ . However an OSS project  $p_3$  with the functionalities of a pure text editor is dissimilar to both  $p_1$  and  $p_2$ , i.e.  $\text{Score}(p_1, p_2)=\text{Score}(p_1, p_3)=1$ .
- An OSS project  $p_4$  which possesses the functionalities of a normal text editor, e.g. those by  $p_3$ , however it embeds the ability to send text files over the network, then it is similar to all the above mentioned project  $p_1$ ,  $p_2$ , and  $p_3$  i.e.  $\text{Score}(p_1, p_4)=3$ ,  $\text{Score}(p_2, p_4)=3$ , and  $\text{Score}(p_3, p_4)=3$ .
- The following apps are considered to be highly similar to each other: Viber, WhatsApp, Messenger since they are all applications that exchange text and multimedia messages over the TCP/IP network.

## **2. How to perform the evaluation**

Each of you will be given 5 queries. For each query, you need to evaluate the results of two different similarity tools, among the four tools mentioned above (MUDABlue, CLAN, RepoPal, and CrossSim). To avoid bias against a specific tool, however you will not know exactly from which tool the query-retrieved projects are produced.

For each query, there are 5 retrieved projects and your task is to evaluate to which degree a pair of <query project, retrieved project> are similar. For each project, there will be a hyperlink to the actual GitHub repository where you can read the **Readme.md** file to better understand the functionalities. In case the Readme.md file is too brief and doesn't provide enough information to make a judgement, please try to investigate the project by consulting its source code files to ascertain its functionalities. It is often the case that you will not be able to understand everything about a project just by looking at its **Readme.md**.

### 3. Similarity Level

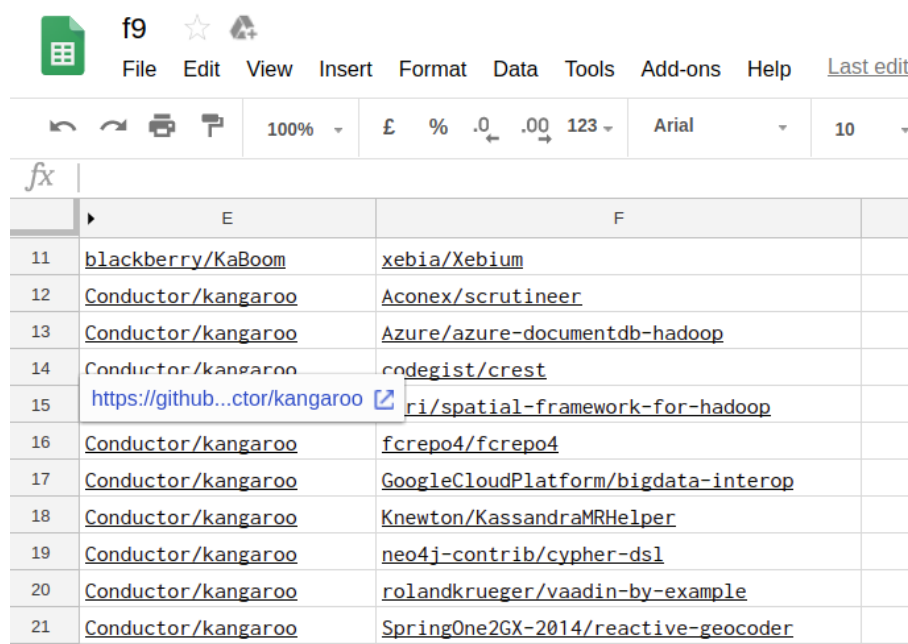
Given a pair of <query project, retrieved project>, you will give a score that specifies the level of similarity between them according to the following descriptions:

Similarity Level	Description	Score
Completely dissimilar	There is absolutely nothing in the retrieved project that is similar to the query project, nothing in it is related to the task and the functionality of the query project.	1
Mostly dissimilar	Only few remotely related requirements are located in the query and retrieved project.	2
Mostly similar	A somewhat large number of implemented requirements are located in the retrieved project that are similar to ones in the query project.	3
Highly similar	The query and the retrieved projects share the same semantic concepts expressed in the task	4

**Table 1.** Similarity scores

### 4. Interface

Each of you will be assigned an Excel sheet where there are five query projects, each corresponds to ten retrieved projects which are the outcomes of **two** similarity tools among the ones mentioned above (MUDABlue, CLAN, RepoPal, and CrossSim). In total, you have to give score for 50 pairs of <query, retrieved> projects. Figure 1 shows an example of an Excel sheet for evaluation. To unhide the link to the actual GitHub project, please hover the mouse on its name.



	E	F
11	<a href="#">blackberry/KaBoom</a>	<a href="#">xebia/Xebium</a>
12	<a href="#">Conductor/kangaroo</a>	<a href="#">Aconex/scrutineer</a>
13	<a href="#">Conductor/kangaroo</a>	<a href="#">Azure/azure-documentdb-hadoop</a>
14	<a href="#">Conductor/kangaroo</a>	<a href="#">codegist/crest</a>
15	<a href="#">https://github...ctor/kangaroo</a>	<a href="#">ri/spatial-framework-for-hadoop</a>
16	<a href="#">Conductor/kangaroo</a>	<a href="#">fcrepo4/fcrepo4</a>
17	<a href="#">Conductor/kangaroo</a>	<a href="#">GoogleCloudPlatform/bigdata-interop</a>
18	<a href="#">Conductor/kangaroo</a>	<a href="#">Knewton/KassandraMRHelper</a>
19	<a href="#">Conductor/kangaroo</a>	<a href="#">neo4j-contrib/cypher-dsl</a>
20	<a href="#">Conductor/kangaroo</a>	<a href="#">rolandkrueger/vaadin-by-example</a>
21	<a href="#">Conductor/kangaroo</a>	<a href="#">SpringOne2GX-2014/reactive-geocoder</a>

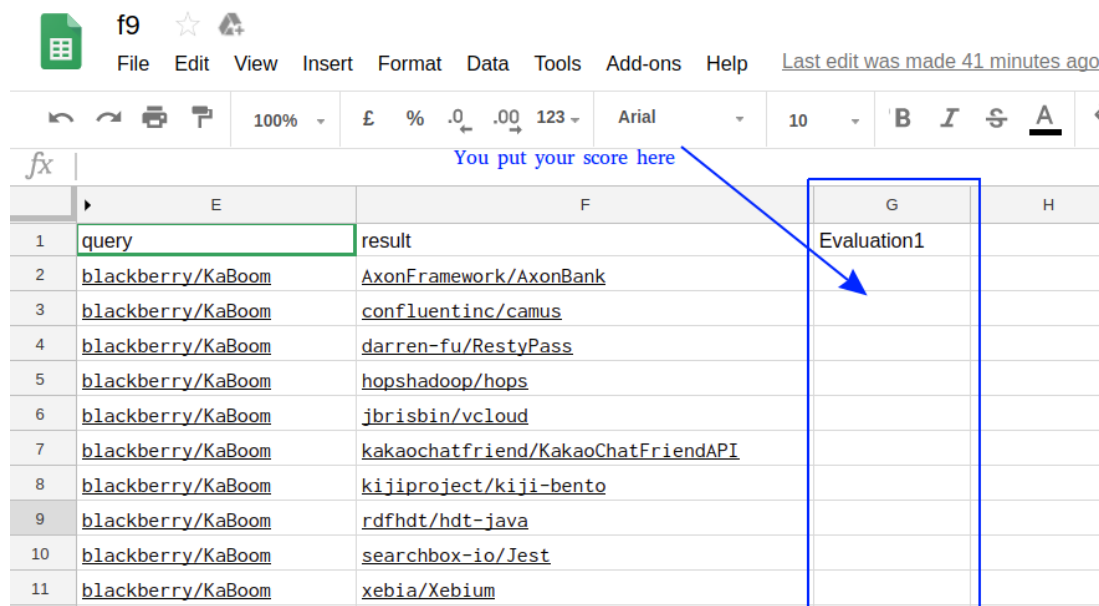
**Figure 1.** Evaluation interface

Sometimes, you may encounter a query project with two retrieved projects that are identical as shown in Figure 2. This is normal since the two tools consider the same project as similar to the query project. In this case, you need to give the same score for both projects.

22	<a href="#">datasalt/pangool</a>	<a href="#">alexholmes/hadoop-utils</a>
23	<a href="#">datasalt/pangool</a>	<a href="#">alexholmes/hadoop-utils</a>
24	<a href="#">datasalt/pangool</a>	<a href="#">alexholmes/hiped2</a>
25	<a href="#">datasalt/pangool</a>	<a href="#">Azure/azure-documentdb-hadoop</a>
26	<a href="#">datasalt/pangool</a>	<a href="#">castagna/jena-grande</a>
27	<a href="#">datasalt/pangool</a>	<a href="#">chubbyjiang/MapReduce</a>
28	<a href="#">datasalt/pangool</a>	<a href="#">hopshadoop/hops</a>
29	<a href="#">datasalt/pangool</a>	<a href="#">laserson/avro2parquet</a>
30	<a href="#">datasalt/pangool</a>	<a href="#">msathis/SQLToNoSQLImporter</a>
31	<a href="#">datasalt/pangool</a>	<a href="#">socrata/datasync</a>

**Figure 2.** A query project with an identical retrieved project

The third column of the sheet is where you put your score (Figure 3). Please note that similarity scores range from 1 to 4 as specified in Table 1, values that are out of this range will not be accepted.



	E	F	G	H
1	query	result	Evaluation1	
2	<a href="#">blackberry/KaBoom</a>	<a href="#">AxonFramework/AxonBank</a>		
3	<a href="#">blackberry/KaBoom</a>	<a href="#">confluentinc/camus</a>		
4	<a href="#">blackberry/KaBoom</a>	<a href="#">darren-fu/RestyPass</a>		
5	<a href="#">blackberry/KaBoom</a>	<a href="#">hopshadoop/hops</a>		
6	<a href="#">blackberry/KaBoom</a>	<a href="#">jbrisbin/vcloud</a>		
7	<a href="#">blackberry/KaBoom</a>	<a href="#">kakaochatfriend/KakaoChatFriendAPI</a>		
8	<a href="#">blackberry/KaBoom</a>	<a href="#">kijiproject/kiji-bento</a>		
9	<a href="#">blackberry/KaBoom</a>	<a href="#">rdfhdt/hdt-java</a>		
10	<a href="#">blackberry/KaBoom</a>	<a href="#">searchbox-io/Jest</a>		
11	<a href="#">blackberry/KaBoom</a>	<a href="#">xebia/Xebium</a>		

**Figure 3.** Put your score in the third column

If you encounter any problem during the evaluation, please do not hesitate to contact us using the information on the first page of this document.

Again, thank you so much for your kind support!

## 5. Examples

The following projects are considered to be similar:

Query project	Retrieved project	Score
<a href="https://github.com/AskNowQA/AutoSPARQL">https://github.com/AskNowQA/AutoSPARQL</a> <b>Description from Readme.md:</b> AutoSPARQL TBSL is a graphical user interface, which allows to answer natural language queries over RDF knowledge bases. It is based on algorithms implemented in the DL-Learner Semantic Web machine learning framework.	<a href="https://github.com/neo4j-contrib/sparql-plugin">https://github.com/neo4j-contrib/sparql-plugin</a> <b>Description from Readme.md:</b> Sparqlify is a scalable SPARQL-SQL rewriter whose development began in April 2011 in the course of the LinkedGeoData project. This tool can create RDF dumps from CSV file based on SML view definitions.	3

Both projects work with SPARQL to query and create RDF data. You may know that RDF is a format for storing Linked Data and SPARQL is the language used to query RDF data. Thus, the two projects share common functionalities. As a results, they are considered to be similar and  $\text{Score}(\text{AskNowQA/AutoSPARQL}, \text{neo4j-contrib/sparql-plugin})=3$ .

The following projects are highly similar:


Query project	Retrieved project	Score
<a href="https://github.com/psaravan/JamsMusicPlayer">https://github.com/psaravan/JamsMusicPlayer</a> <b>Description from Readme.md:</b> Jams is a free, powerful and elegant music player for Android. Jams used to be a trial/paid app on the Play Store.	<a href="https://github.com/TheAndroidMaster/Pasta-Music">https://github.com/TheAndroidMaster/Pasta-Music</a> <b>Description from Readme.md:</b> Pasta Music is a material design music player for android that attempts to create a better user experience than the standard music players which can have too many features or be generally confusing to users. It was created to show an improvement in design and to allow older and slower devices to have quicker access to local music files.	4

Both projects are Android music players so they share common functionalities, such as: open, play music files. Therefore the similarity score is 4 for this pair. We don't care about their internal implementation.

To further elaborate the similarity between software applications concerning their functionalities, we introduce the following examples extracted from Google Play:

### a. Messenger

Similar apps: LINE: Free Calls, WeChat, imo free video, KakaoTalk. These are considered to be similar to Messenger because they are all used for exchanging text and multimedia messages.



## Messenger – Text and Video Chat for Free (Beta)

Facebook Communication ★★★★★ 53,582,348

PEGI 3

Offers in-app purchases

This app is compatible with your device.

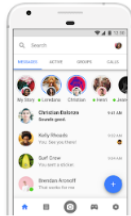


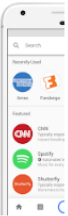
Installed

Message anyone one-on-one or in groups

Capture photos and videos with fun art and effects

Make voice and video calls from anywhere

Connect with that mart

Instantly connect with the people in your life. Messenger is free, fast, and secure.

- Reach anyone. You can use names or phone numbers to find friends.
- Use everywhere. Messenger works across all mobile and desktop devices. You can even connect with people internationally!
- Connect however you want. Send a text message, share a photo, or start a video chat – all in

READ MORE

REVIEWS

Review Policy

Write a Review

### Similar

See more



**LINE: Free Calls**  
LINE Corporation  
Stay in touch with your friends and family using LINE!  
★★★★★



**WeChat**  
WeChat  
The messaging, voice & video call app used by 900 million people  
★★★★★



**imo free video calls**  
imo.im  
Message and video call your family and friends for free!  
★★★★★



**Telegram**  
Telegram Messenger LLP  
Telegram is a messaging app with a focus on speed and security.  
★★★★★




**KakaoTalk: Free**  
Kakao Corporation  
KakaoTalk - the free, fast and fun messenger!  
★★★★★

Facebook See more

## b. Outlook

Similar apps: Email App for Android, Email Exchange, Email-Fast & Secure for Android, Skype for Business, Universal Email. Almost all these apps are used to send Email.



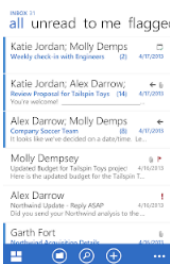
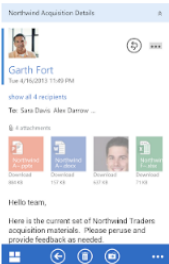
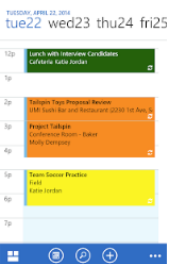
## OWA for Android (Pre-Release)

Microsoft Corporation Communication ★★★★★ 26,465

PEGI 3

This app is compatible with your device.

Add to Wishlist Install

IMPORTANT: Your mailbox must be on the latest version of Office 365 for business (excludes Office 365 Personal and Office 365 Home Premium).

OWA for Android lets you interact with your email, calendar, and contacts from virtually anywhere using your Android phone. You can triage email, manage your schedule, and sync contacts on the go, while protecting your business data.

READ MORE

REVIEWS

Review Policy

2.9

★ 5 7,165

★ 4 3,757

★ 3 3,502

### Similar

See more



**Email App for Android**  
Craigpark Limited  
Quick and easy access to Outlook and Hotmail accounts on the go!  
★★★★★



**Email Exchange**  
Mail Wise  
MailWise - Email without the clutter  
Sync Hotmail, Exchange & more!  
★★★★★



**Email -Fast & Secure**  
Edison Software  
Lightning fast email for Gmail iCloud Yahoo Exchange iMAP AOL  
★★★★★



**Skype for Business**  
Microsoft Corporation  
Skype for Business extends the power of Lync and Skype to your mobile.  
★★★★★



**Universal Email**  
Craigpark Limited  
Designed to support all services - Hotmail, Gmail, Outlook, Yahoo & more!  
★★★★★