**Detecting Similar Repositories on GitHub**

Yun Zhang1, David Lo2, Pavneet Singh Kochhar2, Xin Xia1 , Quanlai Li3, and Jianling Sun1 1College of Computer Science and Technology, Zhejiang University, Hangzhou, China 2School of Information Systems, Singapore Management University, Singapore 3University of California, Berkeley, USA

DOI: [https://xin-xia.github.io/publication/saner17.pdf](https://xin-xia.github.io/publication/saner17.pdf#:~:text=In%20this%20paper%2C%20we%20propose%20a%20novel%20approach,files%29%20which%20are%20not%20considered%20in%20previous%20works.)

Team members: Goian Sergiu-Rares

**Introduction**

MetaSim is a search engine created to retrieve similar (regarding funcitonality and content) GitHub repositories based on metadata and content similarity including readme files, stargazers let by profiles and time interval of stargazers left. The main feature is the use of Repo2Vec, a vector based embedding technique for representing repositories as numerical vectors.

**Installation**

This tool can be accessed very easily by opening the page: <https://metasim-app.github.io/> . It doesn’t require any installation, configuration or any other prerequizites since it can be used just by using google search with „MetaSim repository similarity” and find the react app from the link above.

**Usage**

It was designed to be used as a web application with a very intuitive UI, and it provided all of the functionalities in one page. This page contains some really short and effective instructions, so users have just to search in the text box with github username/repo\_name of the repo for which they want to find other similar repos and select the matching criteria (by name / description / topics / readme). I tried myself for one of my repositories, it takes around 10-15 seconds to fetch all:

A screenshot of a computer

AI-generated content may be incorrect.

Since my repo contains a stock price prediction model that uses LSTM, the tool provided me a lot of repos very similar to mine and not contianing just the LSTM technique, but it has also found repos with RNN, trading bots and other repos that are very similar to my repository’s content. The results may also be filtered by desired columns / values / programming languages and also sorted ascending / descending depending on what column we select

A screenshot of a computer

AI-generated content may be incorrect.

**Conclusion**

This tool is really helpful when you want to search some repositories that you want to perform some research on and it provides great results.