

Browsing NPM packages more effectively with Code Compass

Tom Van Cutsem

 @tvcutsem

Fact: software ecosystems are rapidly expanding



JavaScript



Java



Python



820K+ packages

+454/day



280K+ packages

+275/day

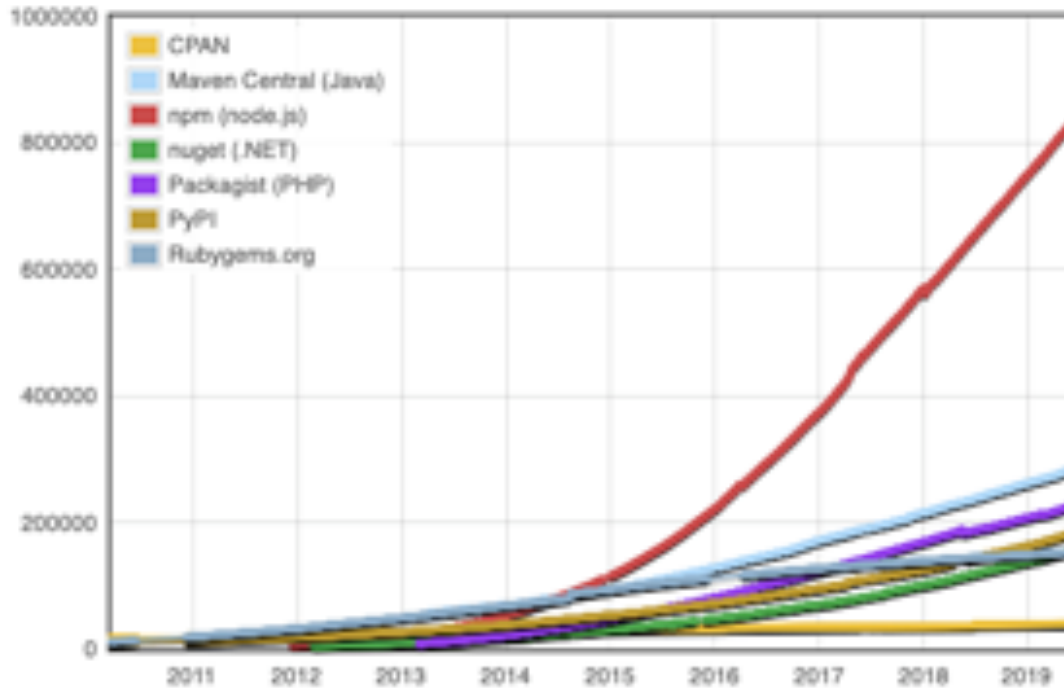


180K+ packages

+130/day

NPM dominates


Module Counts

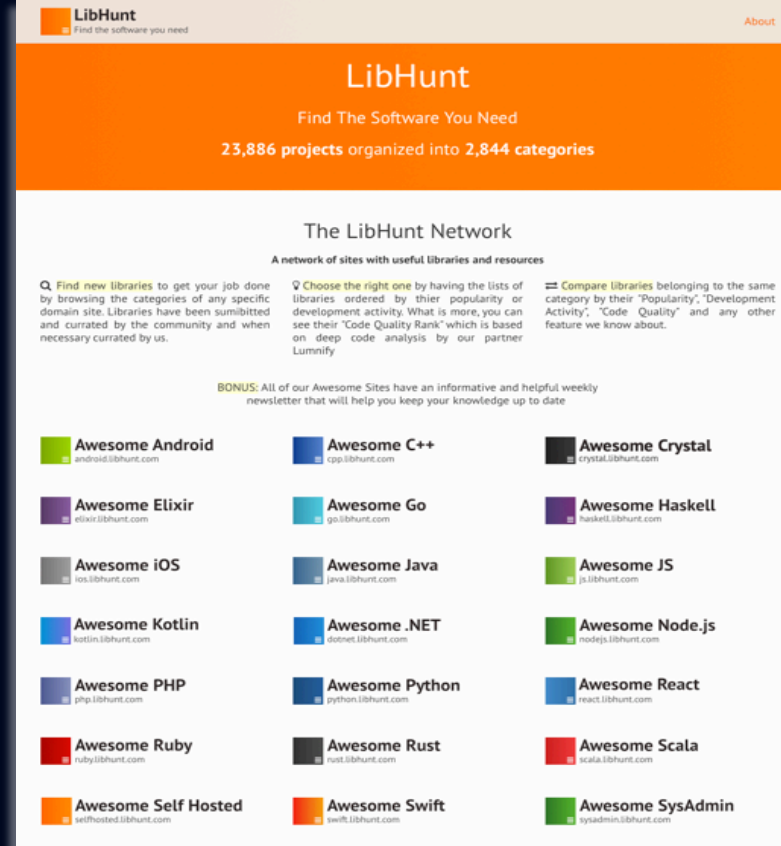


(source: modulecounts.com, June 2019)

That's “awesome”, but...

New problem: how do you find relevant libraries for your development needs?

- Today: manual way of dealing with this
 - “**Awesome**” Lists – community-curated lists of categorized libraries 
 - **LibHunt** – website built on top of awesome lists
 - Indexed 24K libraries
 - Into 3K categories
- But hardly Scalable...
 - Top 6 languages have over **1.5 Million libraries**
 - Only 1.6% is covered by manual indexing efforts



LibHunt
Find The Software You Need
23,886 projects organized into 2,844 categories

The LibHunt Network






















A network of sites with useful libraries and resources

Q Find new libraries to get your job done by browsing the categories of any specific domain site. Libraries have been submitted and curated by the community and when necessary curated by us.

Q Choose the right one by having the lists of libraries ordered by their popularity or development activity. What is more, you can see their "Code Quality Rank" which is based on deep code analysis by our partner Luminy.

Q Compare libraries belonging to the same category by their "Popularity", "Development Activity", "Code Quality" and any other feature we know about.

BONUS: All of our Awesome Sites have an informative and helpful weekly newsletter that will help you keep your knowledge up to date

 Awesome Android android.libhunt.com	 Awesome C++ cpp.libhunt.com	 Awesome Crystal crystal.libhunt.com
 Awesome Elixir elixir.libhunt.com	 Awesome Go go.libhunt.com	 Awesome Haskell haskell.libhunt.com
 Awesome iOS ios.libhunt.com	 Awesome Java java.libhunt.com	 Awesome JS js.libhunt.com
 Awesome Kotlin kotlin.libhunt.com	 Awesome .NET dotnet.libhunt.com	 Awesome Node.js nodejs.libhunt.com
 Awesome PHP php.libhunt.com	 Awesome Python python.libhunt.com	 Awesome React react.libhunt.com
 Awesome Ruby ruby.libhunt.com	 Awesome Rust rust.libhunt.com	 Awesome Scala scala.libhunt.com
 Awesome Self Hosted selfhosted.libhunt.com	 Awesome Swift swift.libhunt.com	 Awesome SysAdmin sysadmin.libhunt.com

Code Compass to the rescue

The screenshot shows the Code Compass web application interface. At the top, there are navigation icons for help, GitHub, and Twitter. The main header features the Code Compass logo and the tagline "The Hitchhiker's Guide to the Software Galaxy". Below this, there are icons for Java, JavaScript, and Python, with the Python icon highlighted. A search bar contains the text "keras" and a button to "Add a Python library to your context". Below the search bar, a note states: "Search results are sourced from PyPI (39135 PYTHON packages indexed to date), enriched with metadata from LibHunt".

Recommended libraries

Filter by tag

- MOST SIMILAR
- Http
- Testing
- Text processing
- Internet
- Scientific
- Engineering
- Utilities
- Www
- Science and data analysis
- Command line tools

tensorflow TensorFlow is an open source machine learning framework for everyone. TensorFlow is an open source software library for high performance numerical computation. Its flexible architecture allows easy deployment of computation across » 290 Apache 2.0	sklearn A set of python modules for machine learning and data mining Use scikit-learn _ instead. » 253	torch Tensors and Dynamic neural networks in Python with strong GPU acceleration » 80
tqdm Fast, Extensible Progress Meter A fast, extensible progress bar for Python and CLI ★ 7.8K » 779 MPLv2.0, MIT Licences	skimage Dummy package that points to scikit-image Image processing in Python ★ 2.6K » 2	joblib Lightweight pipelining: using Python functions as pipeline jobs. Joblib is a set of tools to provide lightweight pipelining in Python. In particular: transparent disk-caching of functions and lazy re-evaluation (memoize pattern) easy simple » 187 BSD
theano	layers	numpy

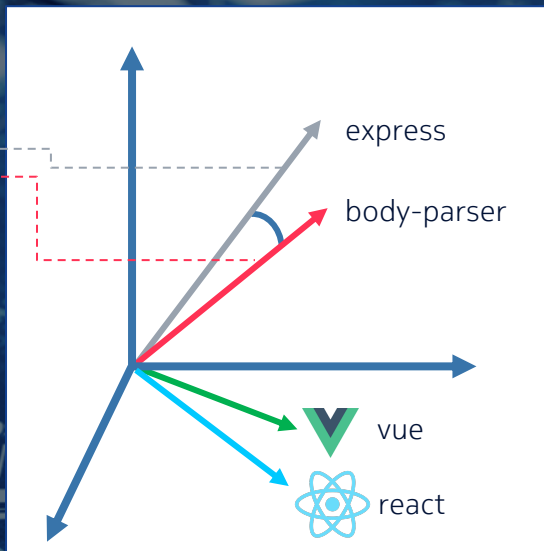
Copyright 2019 Nokia [Terms of Use](#) [Privacy Policy](#) [in](#) [t](#) [f](#) [@](#) [G](#) [d](#) [w](#)

Unsupervised learning from Big Code

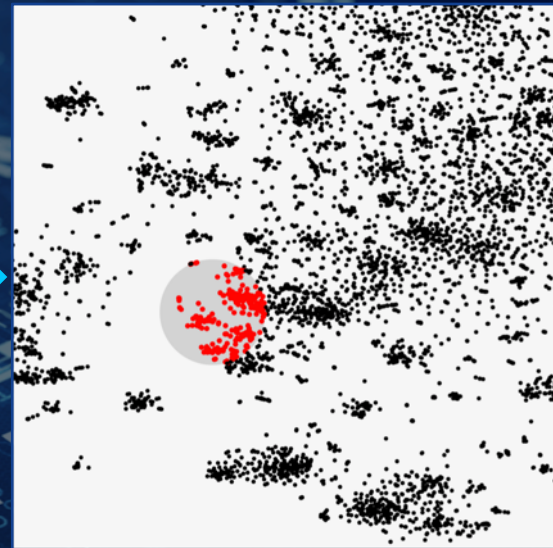
Extract library dependency data from code

```
1 // Dependencies
2 //
3 var express = require('express');
4 var body-parser = require('body-parser');
5 var path = require('path');
6
7 // Sets up the Express App
8 //
9 var app = express();
10 var PORT = 3000;
11
12 // Tells body-parser what type of content to receive
13 app.use(body-parser.json());
14 app.use(body-parser.urlencoded({ extended: true }));
15 app.use(body-parser.text());
16 app.use(body-parser({ type: 'application/vnd.api+json' }));
```

Learn vector representation



Compute nearest neighbors

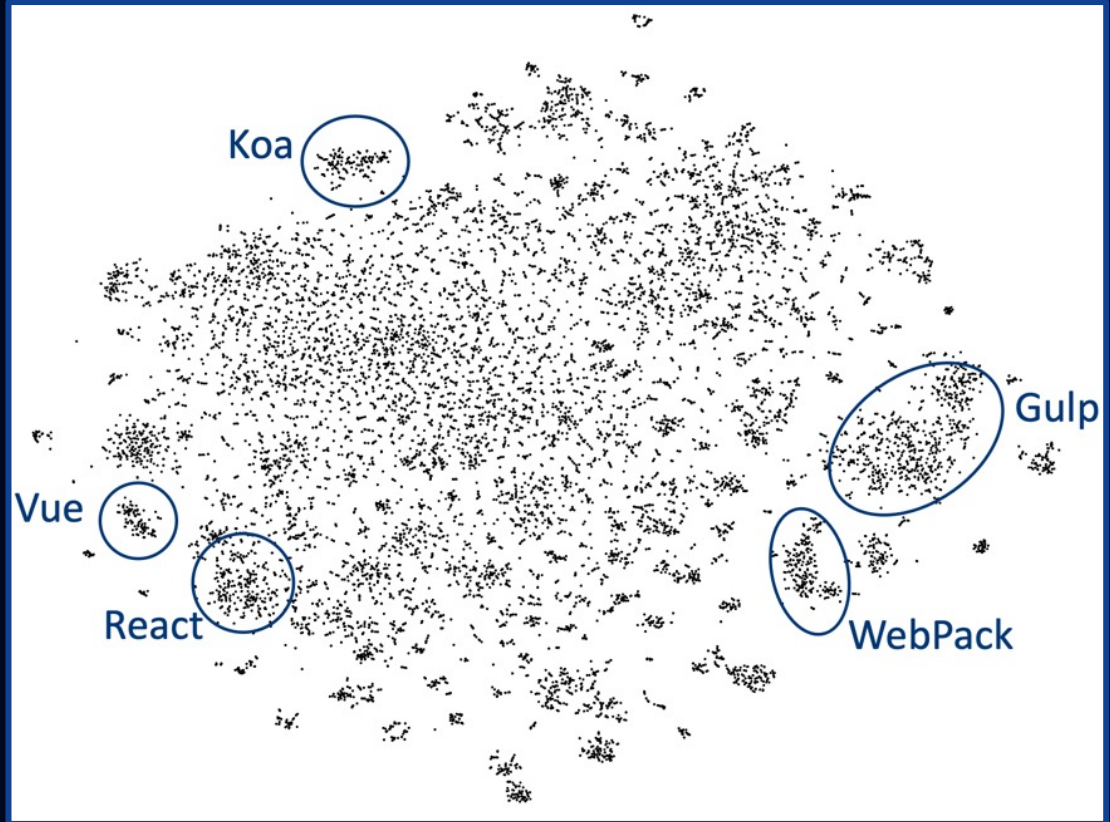
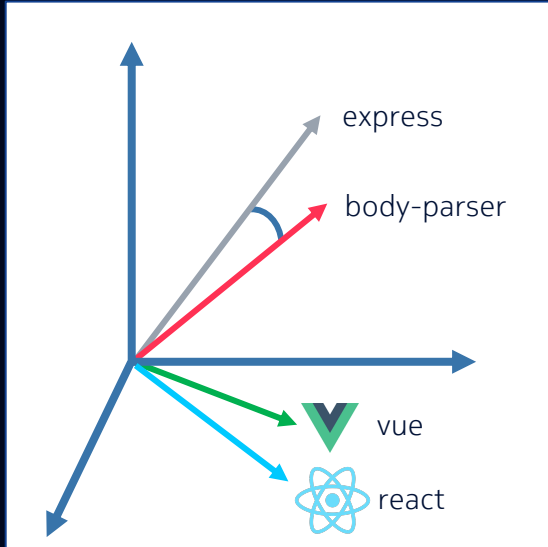


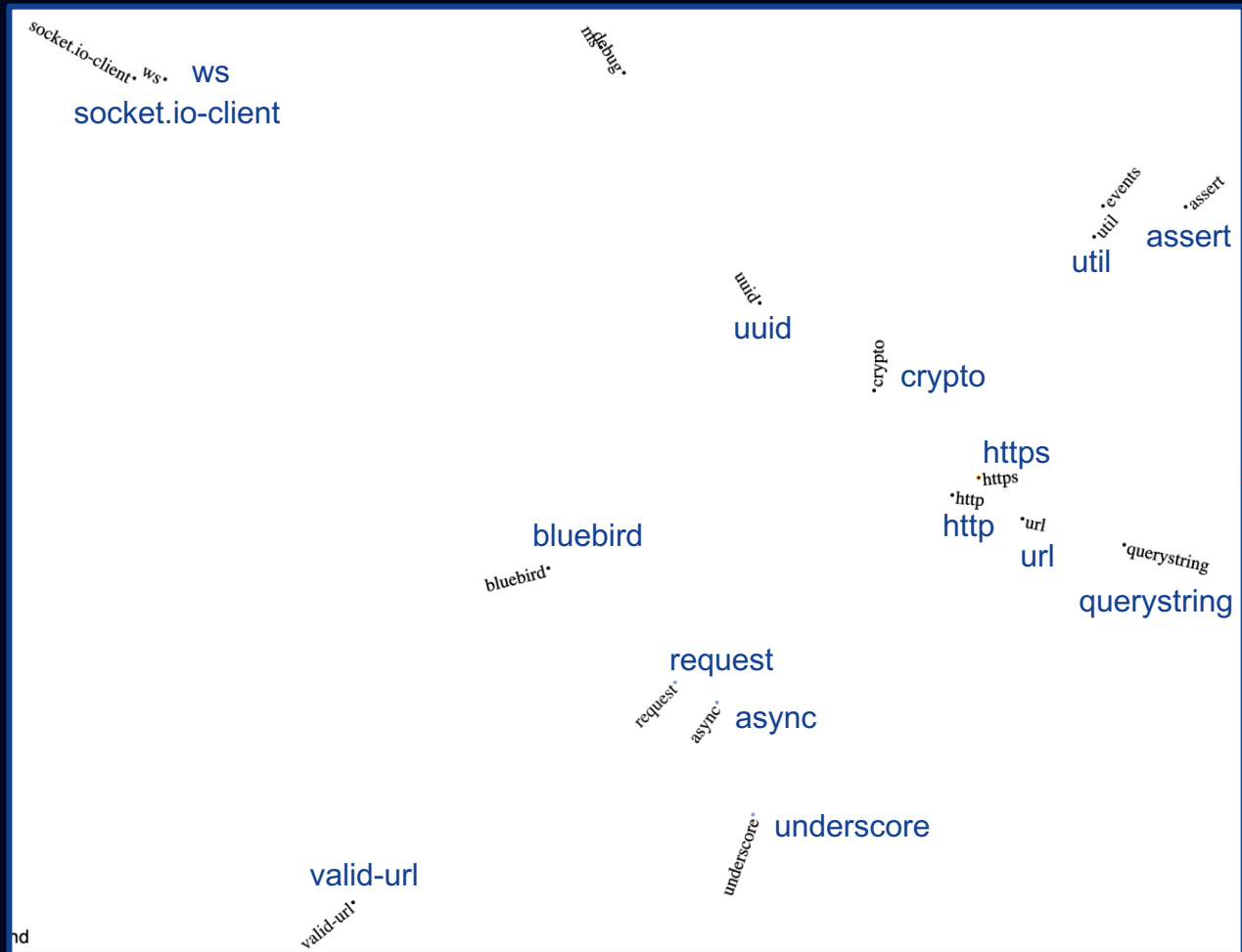
Doesn't Code Compass just recommend popular combo's? **No!**

Search anchor	Code Compass top results	Most Popular combos
<code>mysql</code>	<code>pg</code> (#299) <code>redis</code> (#117) <code>knex</code> (#343) <code>mongodb</code> (#97) <code>nodemailer</code> (#153)	<code>express</code> (#3) <code>body-parser</code> (#13) <code>async</code> (#25) <code>lodash</code> (#12) <code>request</code> (#17)
<code>gm</code> (graphicsmagic)	<code>imagemagick</code> (#1517) <code>sharp</code> (#1040) <code>connect-busboy</code> (#1913) <code>jimp</code> (#1010) <code>canvas</code> (#350)	<code>async</code> (#25) <code>request</code> (#17) <code>express</code> (#3) <code>lodash</code> (#12) <code>crypto</code> (#16)

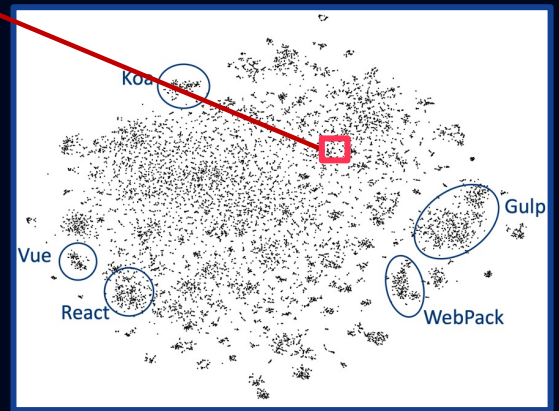
Mapping the JavaScript library ecosystem

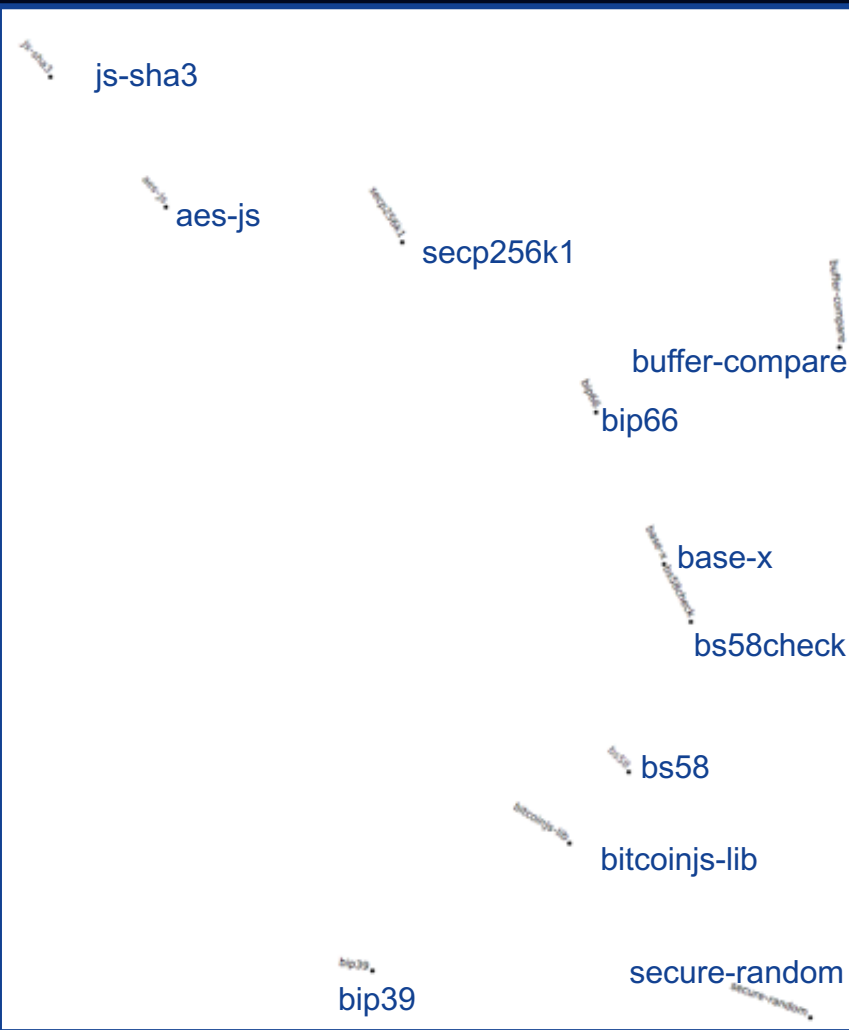
2D projection of a 100D space



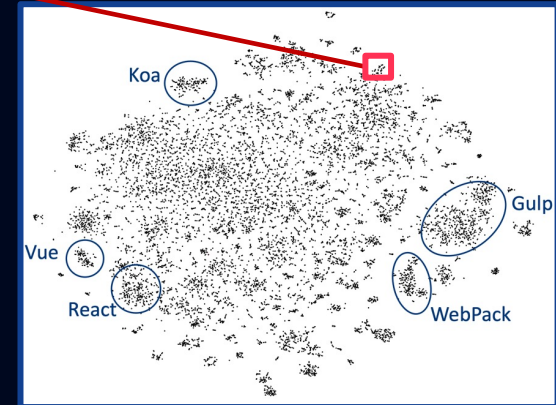


Local neighborhoods reveal meaningful similarities between libraries





Local neighborhoods
reveal meaningful
similarities between
libraries



Data

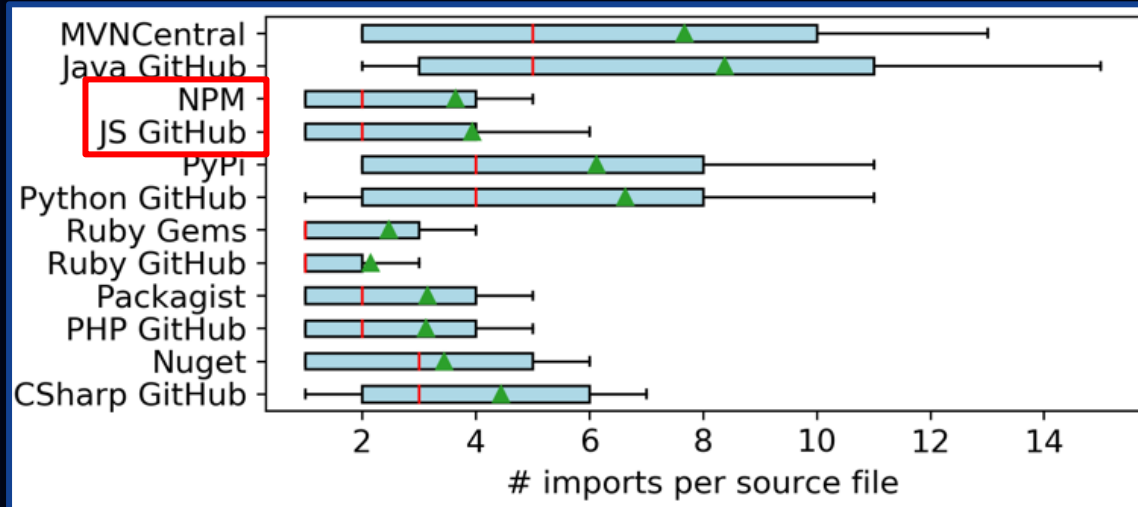
GitHub+NPM JS/TS projects crawled	764K
Source files *.{js, ts} crawled	20.4M
Unique require/import statements crawled	216K
Libraries indexed by import2vec	88.4K



Fun fact: how many modules get imported in a typical JS file?

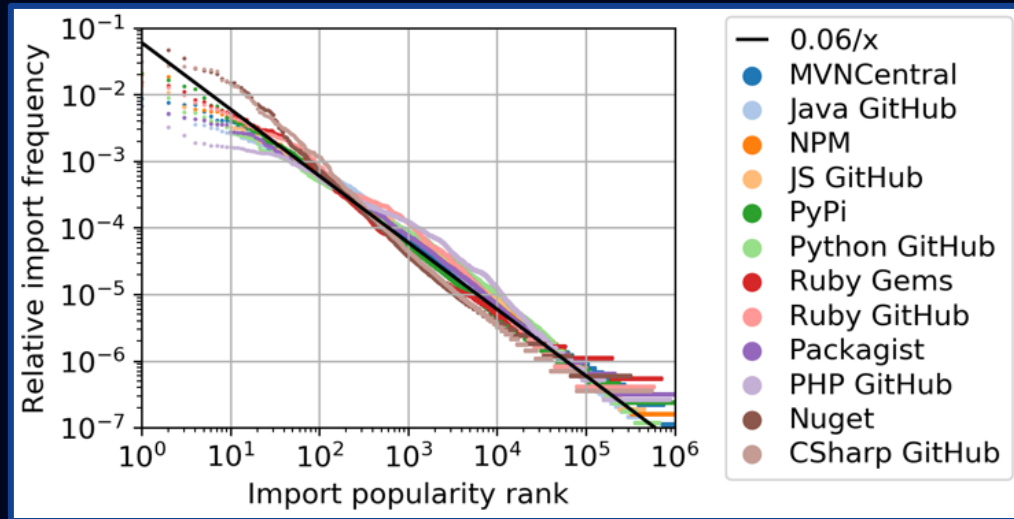
Fun fact: how many modules get imported in a typical JS file?

- Answer: about 4 on average



Fun fact: module imports follow Zipf's Law

- The 2nd most popular module gets imported only half as much as the most popular one
- The n^{th} most popular module gets imported only $\sim 1/n$ as much as the most popular one



Similar to frequency of letters in alphabet, words in text documents, ...

Hunger for more details? Read our paper

Import2vec
Learning Embeddings for Software Libraries

<p>Bart Theeten <i>Nokia Bell Labs</i> Antwerp, Belgium bart.theeten@nokia-bell-labs.com</p>	<p>Frederik Vandeputte <i>Nokia Bell Labs</i> Antwerp, Belgium frederik.vandeputte@nokia-bell-labs.com</p>	<p>Tom Van Cutsem <i>Nokia Bell Labs</i> Antwerp, Belgium tom.van_cutsem@nokia-bell-labs.com</p>
--	--	--

Abstract—We consider the problem of developing suitable learning representations (embeddings) for library packages that capture semantic similarity among libraries. Such representations are known to improve the performance of downstream learning tasks (e.g. classification) or applications such as contextual search and analogical reasoning.

We apply word embedding techniques from natural language processing (NLP) to train embeddings for library packages (“library vectors”). Library vectors represent libraries by similar context of use as determined by import statements present in source code. Experimental results obtained from training such embeddings on three large open source software corpora reveals that library vectors capture semantically meaningful relationships among software libraries, such as the relationship between frameworks and their plug-ins and libraries commonly used together within ecosystems such as big data infrastructure projects (in Java), front-end and back-end web development frameworks (in JavaScript) and data science toolkits (in Python).

Index Terms—machine learning, software engineering, information retrieval

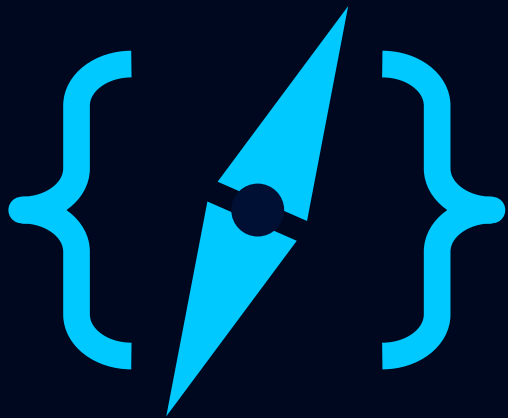
well exceeds over 1 million packages (data as of January 2019) and so most packages in the “long tail” remain undiscoverable through manual curation.

The size and scale of today’s software ecosystems suggests that a machine learning approach could help us build tools that help developers more effectively navigate them. However, for most learning algorithms to be applied successfully to this problem, we require a mathematical representation of libraries, preferably one that represents similar libraries by similar representations.

This paper addresses the question whether we can leverage techniques from natural language processing, in particular word embeddings, to learn meaningful distributed representations of software libraries from large codebases. Just like word embeddings learn to represent similar words by similar dense vector representations based on the words’ similar context of use, we aim to learn a dense vector representation of libraries

arXiv:1904.03990 [cs.SE] 27 Mar 2019

<https://arxiv.org/abs/1904.03990>
(google “import2vec”)



Code Compass

Contextual search for code

Give Code Compass a try. Thanks!



bell-labs.com/code-compass



github.com/nokia/code-compass



[@tvcutsem](https://twitter.com/tvcutsem)



NOKIA