Generic Search and Health Analysis on Github repos

pranjal gupta 2013B4A7470P
a anjana 2013A3A7244P
shubhi rastogi 2013B3A7521P
himanshu s dhoni 2013A7PS187P

Introduction

GitHub is the codenetwork for developers. Anyone who's worth it's salt is there. GitHub is already a powerful platform and will surely improve in the coming years. Due to network effects and the intertwining between projects Github has the highest rate of lock in, and more and more third parties directly integrate into the Github API.

But we believe that the number of open source projects is set to increase and GitHub has a unique opportunity to position itself ,not only as a community, but as a development workbench for distributed teams.

The use and relevance of this online platform to developers greatly depends on the ability of it to return context-based search results to queries.What we aim to do via the project is to create a generic priority based search engine for GitHub repositories.

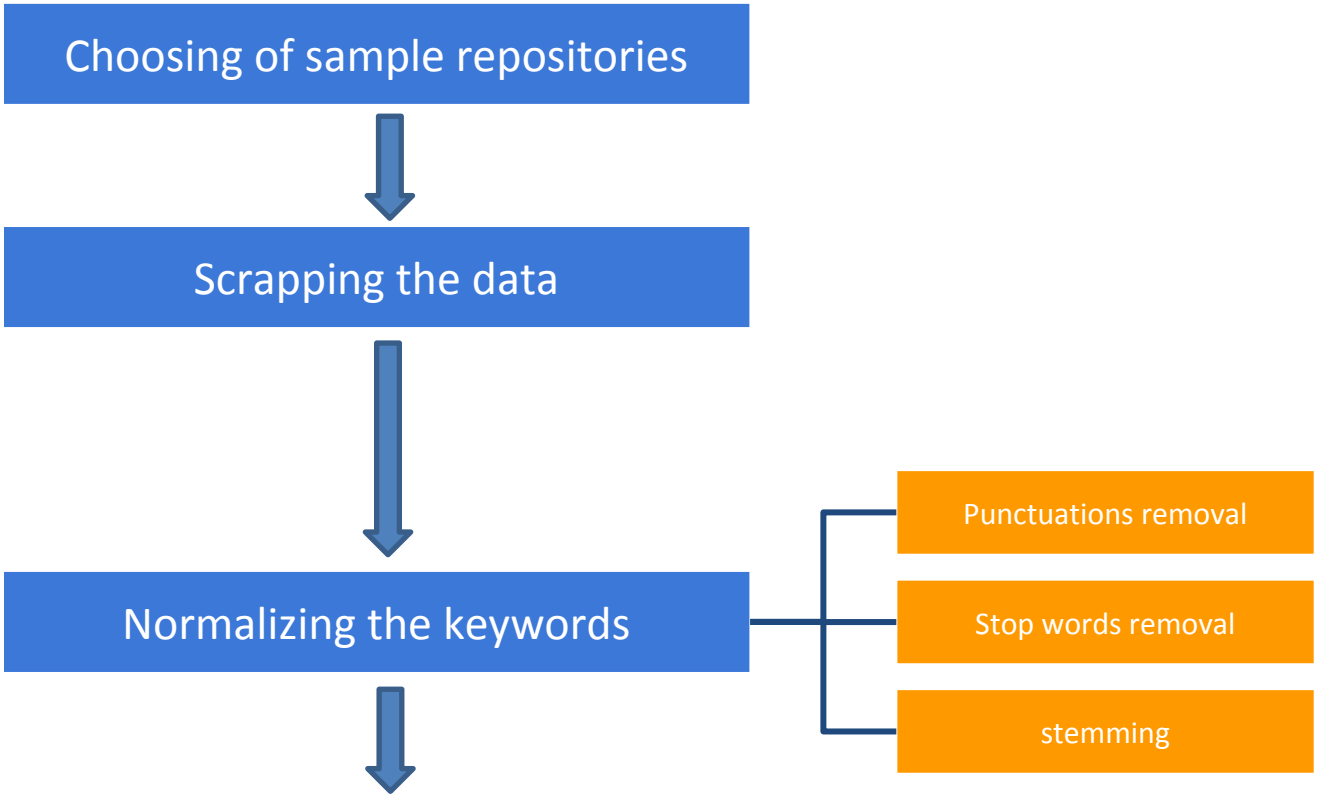
Why is it different?

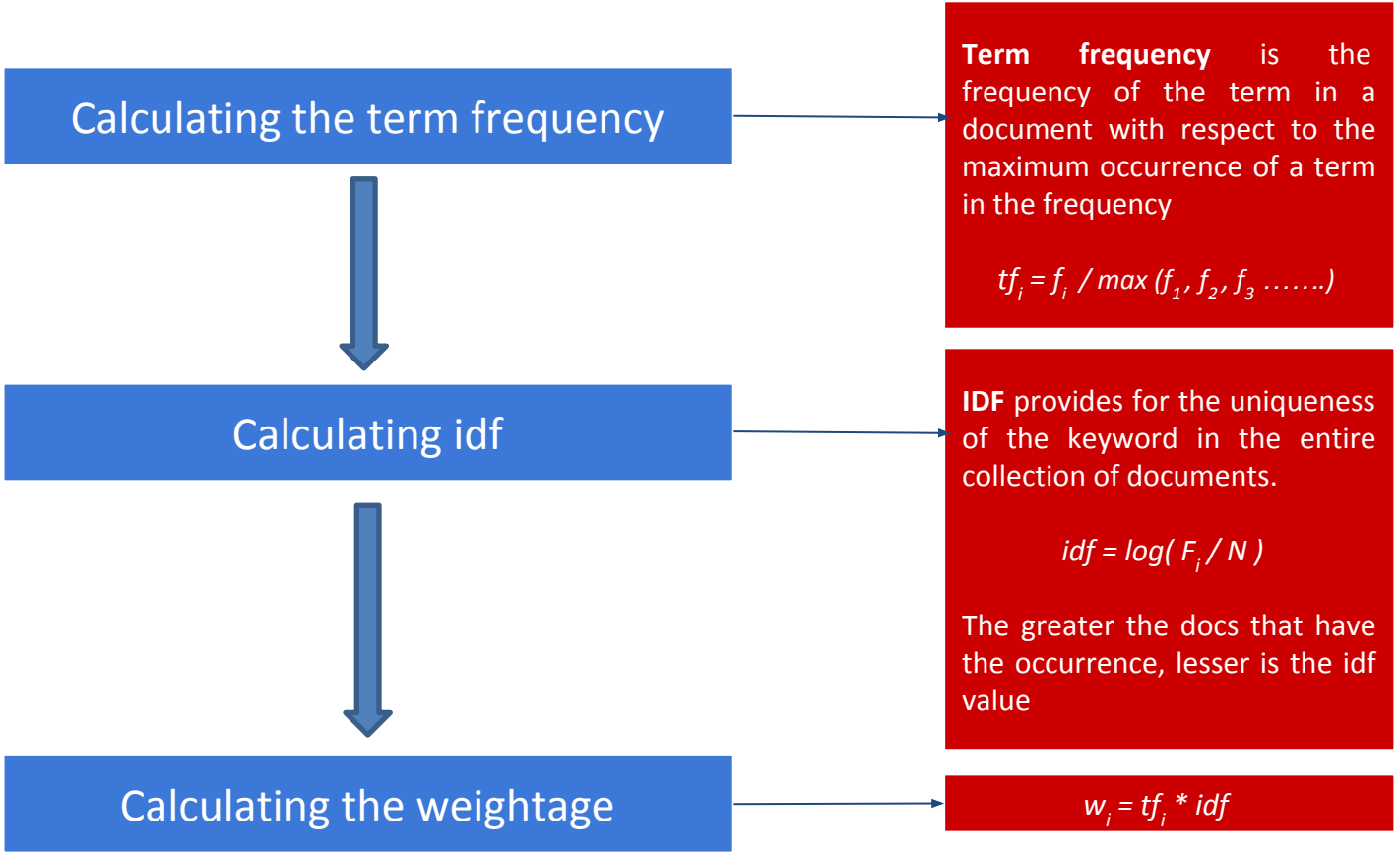
Present scenario on GitHub

- The present search criteria and algorithm used doesn't check for the content of the repositories for keywords, rather focus exclusively on the title and description
- Ranking of the occurrence is not calculated (Boolean Model of IR). Weightage based search is not implemented
- **Popularity** and **stability** of a repo is not considered while placing results.

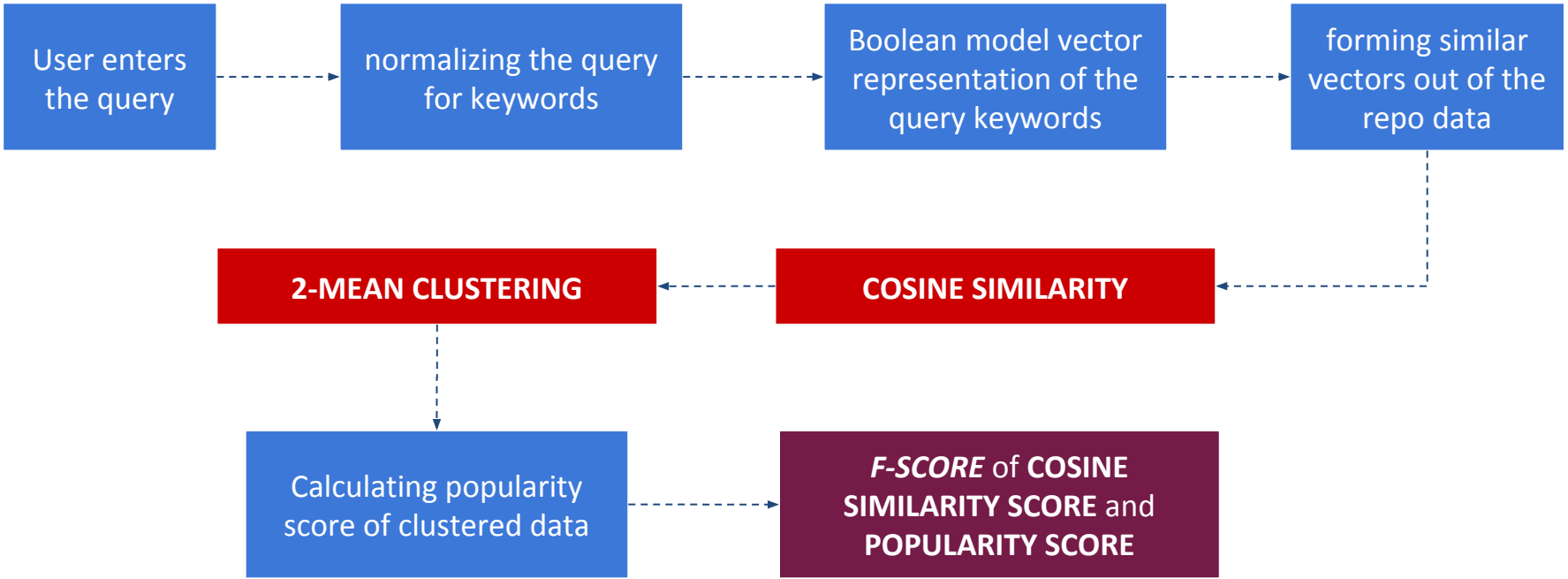
What we propose is a search algorithm that removes all of these shortcomings resulting in useful and priority based results turning up for queries. This will connect the end user with more accurate results.

Completed work plan

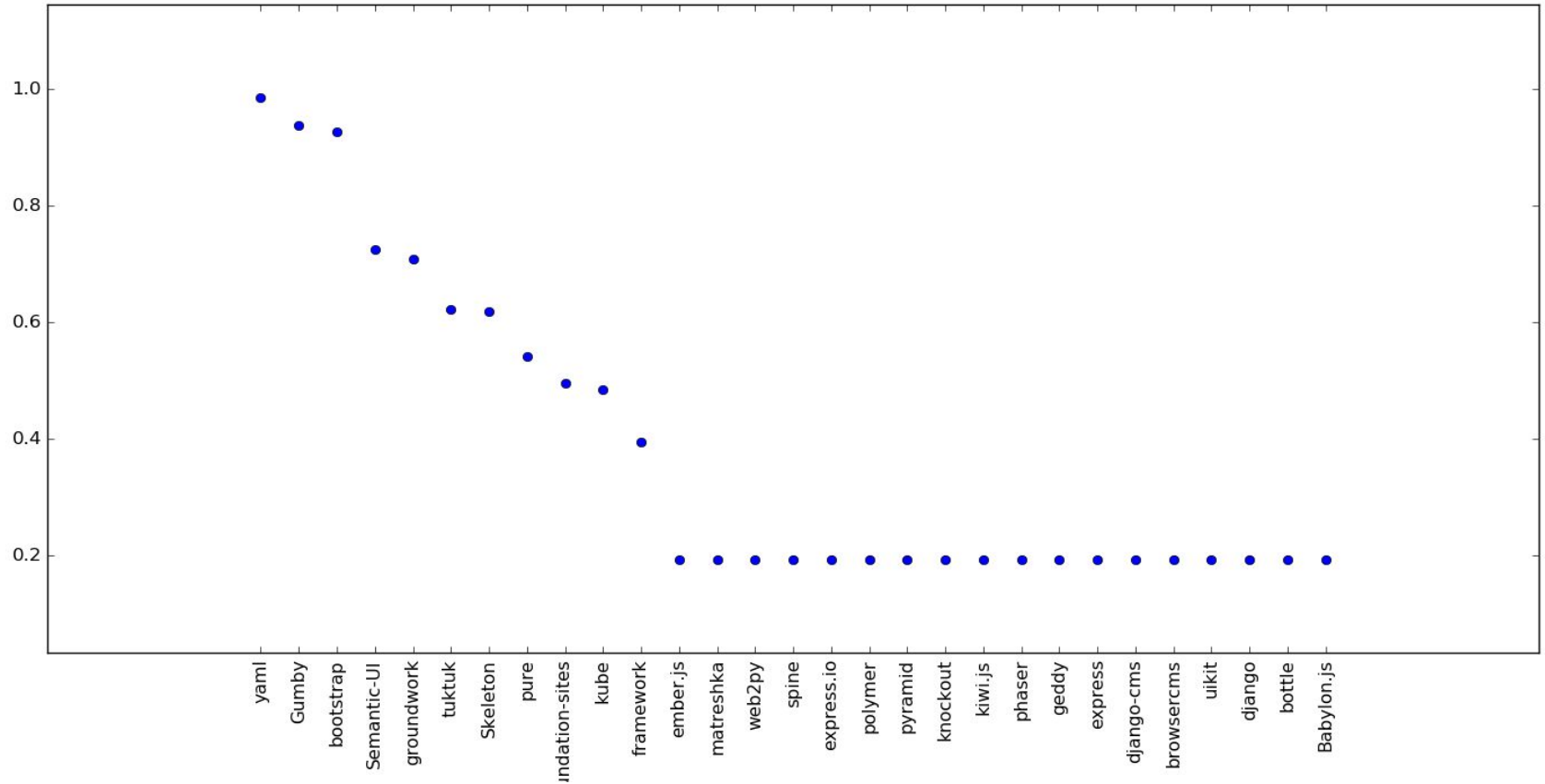



Calculating the term frequency
Term frequency is the frequency of the term in a document with respect to the maximum occurrence of a term in the frequency
$tf_i = f_i / max (f_1, f_2, f_3 \ldots\ldots)$
Calculating idf
IDF provides for the uniqueness of the keyword in the entire collection of documents.
$idf = log(F_i / N)$
The greater the docs that have the occurrence, lesser is the idf value
Calculating the weightage
$w_i = tf_i * idf$

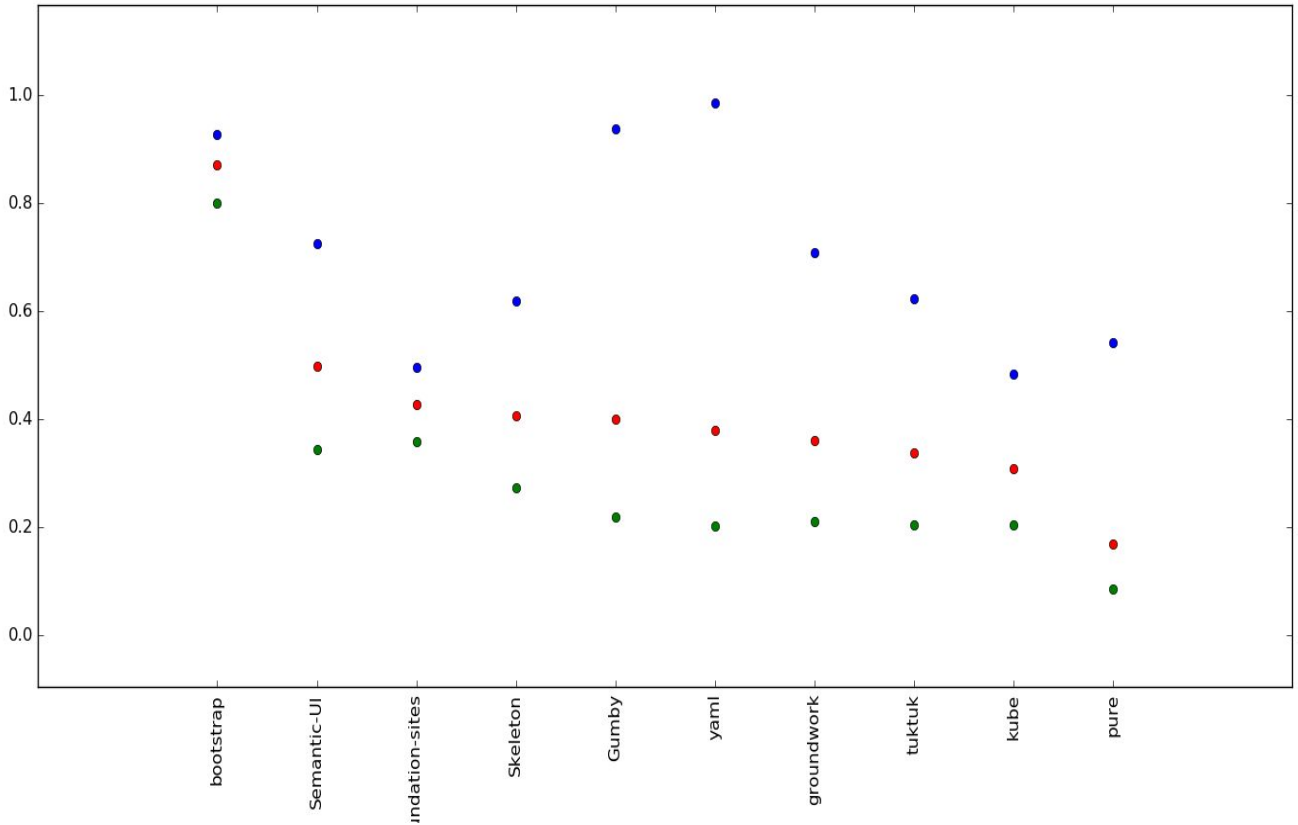
Query processing



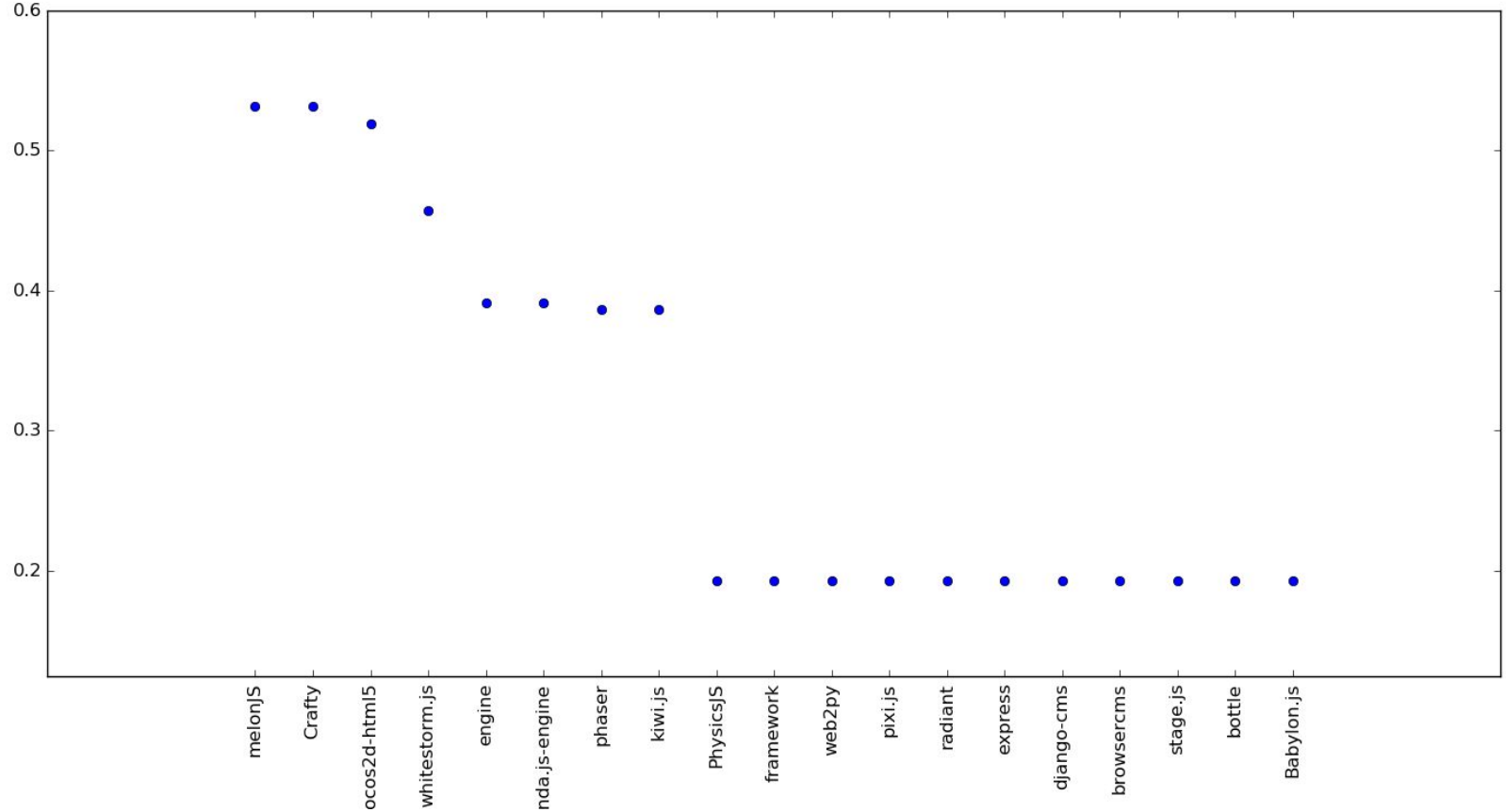
query : “responsive css frameworks”



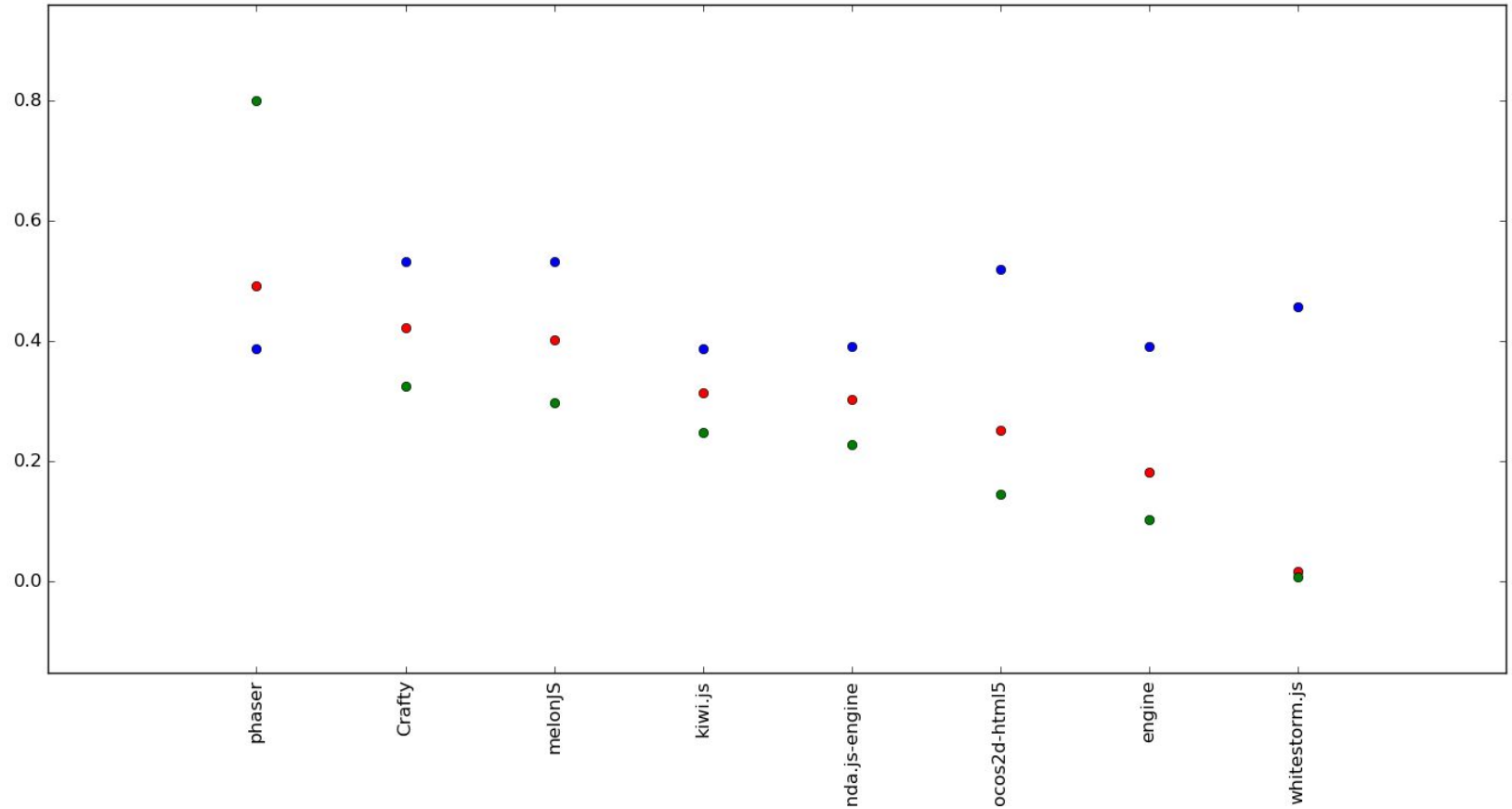
query : “responsive css frameworks”



query : “webgl game engines”



query : “responsive css frameworks”



To be incorporated

- The stability of a repository in GitHub depends on the closing of **issues** by the contributors that are raised by the particular users of the repository. This data needs to be incorporated in the result
- **Pull requests**
- **Milestones** have to be given appropriate accountance while discussing relevance
- Development of user interface

thank you