Search Parasparam 2021

Parasparam 2021

View All Results ()

Notifications

Mark All as Read

Action Items

View All

V

Profile

Log Out

HOME

FAQS

MESSAGES

NEWS

Subscribe

HOME

FAQS

MESSAGES

NEWS

1

Duplicate defect finder using text mining (D2677)

Subscribe

Team - divya nambiar Submitted: Jun 7 2021

Category: Big Data and Analytics

Status: Submitted

comparison defect text analysis

Description Attachments (1) Votes

Request To Join Team

Find Teammates

Submission Tea 1 + **Add/Edit**

D V

Details

Vote Score:	1
Votes:	1
Rank:	5
Unique Views:	2
Total Views:	14
Comments:	0
Favorited:	0

Linked Submissions:

No Linked Submissions

Duplicates defects are extra overhead for an 'as a service' Company like HPE, as we have frequent releases.

Also time, cost and effort of managing duplicates are mainly redundant. Also because of various ways to describe same defect, it is hard to investigate duplicate defect manually.

Ideally, an efficient tool should prevent the analysis and creation of duplicate defect. This tool also can list the top N similar defect and allow the developers and testers how much percentage it is similar to the given defect.

Also It helps developers especially new joiners to get the context of what piece of codebase to look at, what DB queries to do and in general get an idea of what needs to be debugged by looking at old completed similar defects.

The similar applicable to Epics and stories as well. This demand is the motivation for us to design and develop a search based duplicate defect finder using text mining.

Expertise Required

Text Mining

What is the problem statement for your project?

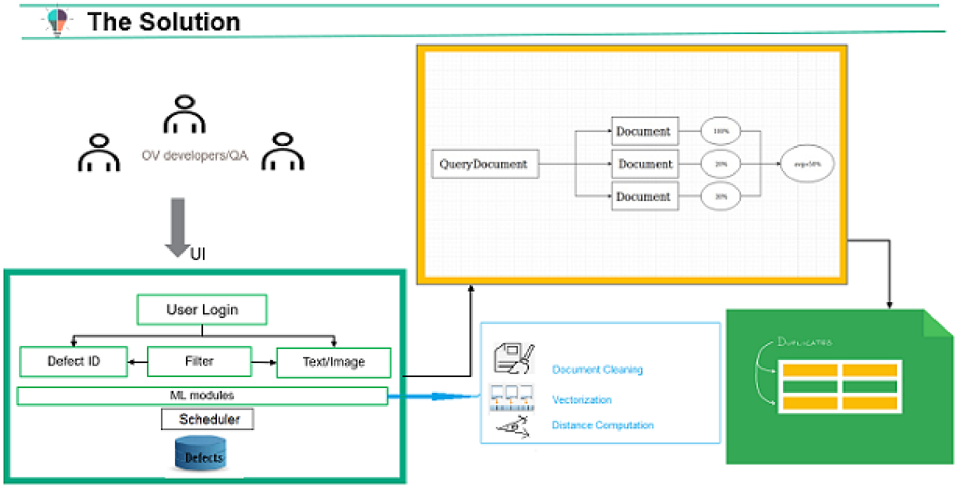
A defect can be reported at any stage during the software development or post release. Preventing the introduction of duplicate defect in the first place is important, as it impact the quality of product from customer point of view.

Normally the tester or developer will not track the existing problem as the time and effort is more there. Ideally, a defect tracking system can prevent users from reporting duplicate defects with minimal overhead.

For an incoming defect report, the system should be able to detect and/or filter duplicates and hence offer the defect triages a list of top-N similar reports from the defect repository.

This allows triages to compare the similarity of the incoming defect report with the suggested duplicates. The same can be done for stories and epics as well.

Please describe the proposed solution



In the proposed solution a scheduler will query all the defect, including the attachments, based on the search query, which we need to compared against from the bug tracker like Jira, converting them into a JSON of vector and saving them locally. There is a login for each user so that they can customise their search query and save that. This involve the below steps:

1) Document cleaning: Remove words which are irrelevant for text analysis. Mainly includes removing stop words and converting all word to lower etc. We will consider the text from attachments as well, which are extracted by python image processor.

2) Transform the documents into a vector of numbers using TF/IDF (Term Frequency vs Inverse Document Frequency).

3) Distance Computation: Compute the cosine similarity between the query document vector and other documents. The cosine (dot product) of the same vectors is 1, dissimilar/perpendicular ones are 0, so the dot product of two vector-documents is some value between 0 and 1, which is the measure of similarity amongst them. And List the top N similar defect based on the cosine similarity. User can set the threshold, so that only the defect similarity > threshold will be displayed.

Provide the evidence that the proposed solution works

The tool is already deployed and used by HPE-OneView developers and testers for finding and tracking similar defect s. Also we had got an enhancement request for stories/customer found incidences which has been already incorporated in the tool.

Provide details of competitive approaches

Issue/Bug trackers like Jira has search option which will compare exact match of texts. There is no software which will compare defect using text analysis. Also bug tracker search will not compare the test in images/screenshots, emails, log snippets etc which will tell alot about the defect. But our soultion convert image to tests using AI libraries like PIL and inclede those also for the comparison, which will give more accuracy.

Current Status

The Similar Defect Finder is currently integrated with Jira and POC for same is available here -> http://15.212.160.181:443

Next Steps

1) Enhance and Integrate the solution for other issue/bug types and defect tracking tools apart from Jira.

2) Add capability to create new defects from tool itself in case there are no existing duplicates found.

3) Add support for stories and epics.

Acknowledgements

References

https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/

https://www.atlassian.com/software/jira/guides/expand-jira/jql

https://towardsdatascience.com/the-best-document-similarity-algorithm-in-2020-a-beginners-guide-a01b9ef8cf05

Comments (0)

No comments